Hindawi Publishing Corporation
Advances in Decision Sciences
Volume 2012, Article ID 728980, 25 pages
doi:10.1155/2012/728980

*Research Article*

# Solving the Omitted Variables Problem of Regression Analysis Using the Relative Vertical Position of Observations

**Jonathan E. Leightner[1] and Tomoo Inoue[2]**

[1] *Hull College of Business Administration, Augusta State University, 2500 Walton Way, Augusta, GA 30904, USA*
[2] *Faculty of Economics, Seikei University, 3-3-1 Kichijoji-kitamachi, Musashino-shi, Tokyo 180-8633, Japan*

Correspondence should be addressed to Jonathan E. Leightner, jleightn@aug.edu

Received 9 April 2012; Accepted 11 October 2012

Academic Editor: David Bulger

Copyright © 2012 J. E. Leightner and T. Inoue. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The omitted variables problem is one of regression analysis' most serious problems. The standard approach to the omitted variables problem is to find instruments, or proxies, for the omitted variables, but this approach makes strong assumptions that are rarely met in practice. This paper introduces best projection reiterative truncated projected least squares (BP-RTPLS), the third generation of a technique that solves the omitted variables problem without using proxies or instruments. This paper presents a theoretical argument that BP-RTPLS produces unbiased reduced form estimates when there are omitted variables. This paper also provides simulation evidence that shows OLS produces between 250% and 2450% more errors than BP-RTPLS when there are omitted variables and when measurement and round-off error is 1 percent or less. In an example, the government spending multiplier, $\partial GDP/\partial G$, is estimated using annual data for the USA between 1929 and 2010.

## 1. Introduction

One of regression analysis' most serious problems occurs when omitted variables affect the relationship between the dependent variable and included explanatory variables.[1] If researchers estimate without considering that the true slope, $\beta_1$, is affected by other variables, then they obtain a slope estimate that is a constant,[2] in contrast to the true slope which varies with $q$. In this case the regression coefficients are hopelessly biased and all statistics are inaccurate ($X'e \neq 0$):

$$Y = \alpha_0 + \beta_1 X, \tag{1.1}$$

$$\beta_1 = \alpha_1 + \alpha_2 q^m, \tag{1.2}$$

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 X q^m. \tag{1.3}$$

By substituting (1.2) into (1.1) to produce (1.3), we can see that an easy way to model this omitted variables problem is to use an interaction term, $\alpha_2 X q^m$, which is what we do for the remainder of this paper. However, it is important to realize that this modeling approach captures a much more general problem—a problem that occurs any time omitted variables affect the true slope.

The standard approach to dealing with the omitted variables problem is to use instrumental variables or proxies. However, to correctly use these approaches, the researcher must know how to correctly model the omitted variable's influence on the dependent variable and the relationship between the instruments and the omitted variables. These requirements are often impossible to meet as many researchers do not even know what important variables they are omitting, much less how to correctly model their influence on the dependent variables via proxies.[3] One implication of Kevin Clarke's papers [1, 2] is that including additional proxies may increase or decrease the bias of the estimated coefficients. The approach taken in this paper avoids the problems discussed by Clarke by directly using the combined effects of all omitted variables instead of trying to replace individual omitted variables.

Specifically, this paper introduces the third generation of a technique which produces reduced form estimates of $\partial Y/\partial X$, which vary from observation to observation due to the influence of omitted variables, without using instruments and, thus, without having to make the strong assumptions required by instrumental variables. In essence, this technique recognizes that (for all observations associated with a given value for the known independent variable) the vertically highest observations will be associated with values for the omitted variables that increase $Y$ the most and that the observations on the bottom will be associated with omitted variable values that increase $Y$ the least.

Section 2 of this paper provides an intuitive explanation of this new technique, named "best projection reiterative truncated projected least squares" (BP-RTPLS), and provides a very brief survey of the literature concerning the predecessors to BP-RTPLS. Section 3 presents a theoretical argument that BP-RTPLS estimates will be unbiased. Section 4 presents simulation results that show that ordinary least squares (OLS) produce error that is between 250% and 2450% of the error of BP-RTPLS when there is 1 percent measurement/round-off error, when sample sizes of 100 or 500 observations are used, and when the omitted variable makes a 10 percent, 100 percent, or 1000 percent difference to the true slope. Section 5 provides an example, and Section 6 concludes.

## 2. An Intuitive Explanation of BP-RTPLS and Literature Survey

The key to understanding BP-RTPLS is Figure 1. To construct Figure 1, we generated two series of random numbers, $X$ and $q$, which ranged from 0 to 100. We then defined

$$Y = 100 + 10X + 0.4qX. \tag{2.1}$$

Thus the true value for $\partial Y/\partial X$ equals $10+0.4q$. Since $q$ ranges from 0 to 100, the true slope will range from 10 (when $q = 0$) to 50 (when $q = 100$). Thus $q$ makes a 500 percent difference to the slope. In Figure 1, we identified each point with that observation's value for $q$. Notice that the upper edge of the data corresponds to relatively large $qs - 92, 98, 98$, and 95. The lower edge of the data corresponds to relatively small $qs - 1, 1, 1, 1$, and 6. This makes sense since as $q$ increases so does $Y$, for any given $X$. For example, when $X = 85$, reading the values of

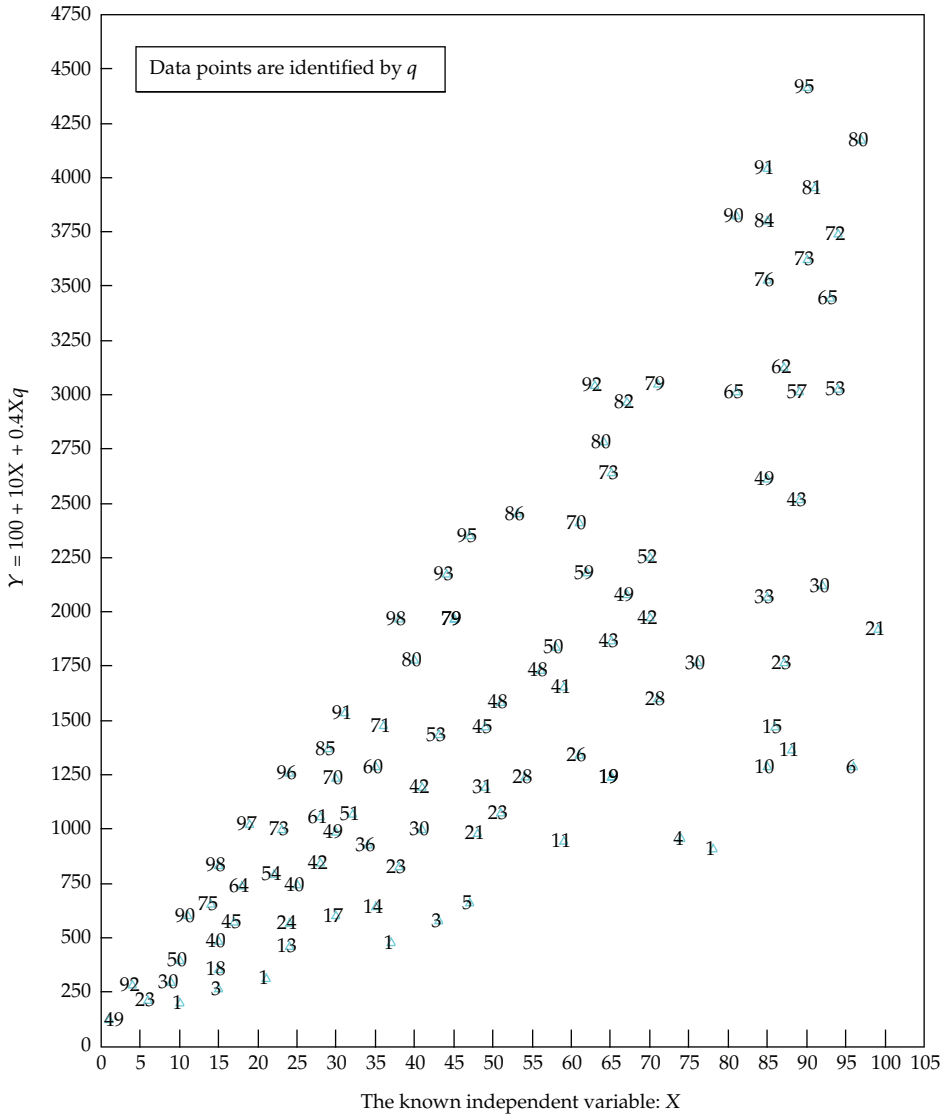



**Figure 1:** The intuition behind 4D-RTPLS.

$q$ from top to bottom produces 91, 84, 76, 49, 33, and 10. Thus the relative vertical position of each observation is directly related to the values of $q$.[4]

An alternative way to view Figure 1 is to realize that, since the true value for $\partial Y/\partial X$ equals $10 + 0.4q$, the slope, $\partial Y/\partial X$, will be at its greatest value along the upper edge of the data where $q$ is largest and the slope will be at its smallest value along the bottom edge of the data where $q$ is smallest. This implies that the relative vertical position of each observation, for any given $X$, is directly related to the true slope.

Now imagine that we do not know what $q$ is and that we have to omit it from our analysis. In this case, OLS produces the following estimated equation: $Y = 87.3 + 30.13X$ with an $R$-squared of 0.6065 and a standard error of the slope of 2.452. On the surface, this OLS regression looks successful, but it is not. Remember that the true equation is $Y = 100 + 10X + 0.4qX$. Since $q$ ranges from 0 to 100, the true slope (true derivative) ranges from 10 to 50 and

OLS produced a constant slope of 30. OLS did the best it could, given its assumption of a constant slope-OLS produced a slope estimate of approximately $10 + 0.4E(q) = 10 + 0.4(50) = 30$. However, OLS is hopelessly biased by its assumption of a constant slope when, in truth, the slope is varying.

Although OLS is hopelessly biased when there are omitted variables that interact with the included variables, Figure 1 provides us with a very important insight—even when we do not know what the omitted variables are, even when we have no clue how to model the omitted variables or measure them, and even when there are no proxies for the omitted variables, Figure 1 shows us that the relative vertical position of each observation contains information about the combined influence of all omitted variables on the true slope. BP-RTPLS exploits this insight. We will first explain 4D-RTPLS (Four Directional RTPLS), UD-RTPLS (Up Down RTPLS), and LR-RTPLS (Left Right RTPLS). BP-RTPLS is the best estimate produced by 4D-RTPLS, UD-RTPLS, and LR-RTPLS.

4D-RTPLS begins with a procedure similar to two stage least squares (2SLS). 2SLS is used to eliminate simultaneous equation bias. In the first stage of 2SLS, all right hand side endogenous variables are regressed by all exogenous variables. The data are plugged into the resulting equations to create instruments for the right hand side endogenous variables. These instruments are then used in the second stage regression. The first stage procedure cuts off and discards all the variation in the right hand side endogenous variables that is not correlated with the exogenous variables.

In a similar fashion, 4D-RTPLS draws a frontier around the top data points in Figure 1. It then projects all the data vertically up to this frontier. By projecting the data to the frontier, all the data would correspond to the largest values for $q$. However, there is a possibility that some of the observations will be projected to an upper right hand side horizontal section of the frontier. For example, the 80 which is closest to the upper right hand corner of Figure 1 would be projected to a horizontal section of the frontier. This horizontal section does not show the true relationship between $X$ and $Y$, and it needs to be eliminated (truncated) before a second stage regression is run through the projected data. This second stage regression (OLS) finds a truncated projected least squares (TPLS) slope estimate for when $q$ is at its most favorable level and this TPLS slope estimate is then appended to the data for the observations that determined the frontier.

The observations that determined the frontier are then eliminated and the procedure repeated. We can visualize this removal as "peeling away" the upper frontier of the data points. As the process is iterated, we peel away the data in successive layers, working downward through the set of data points. The first iteration finds a TPLS slope estimate when the omitted variables cause $Y$ to be at its highest level, ceteris paribus. The second iteration finds a TPLS slope estimate when the omitted variables cause $Y$ to be at its second highest level, and so forth. This process is stopped when an additional regression would use fewer than ten observations (the remaining observations will be located at the bottom of the data). It is important to realize that the omitted variable, $q$, in this process will represent the combined influence of all forces that are omitted from the analysis. For example, if there are 1000 forces that are omitted where 600 of them are positively related to $Y$ and 400 are negatively related to $Y$, then the first iteration will capture the effect of the 600 variables being at their largest possible levels and the 400 being at their lowest possible levels.

Just as the entire dataset can be peeled down from the top, the entire dataset also can be peeled up from the bottom. Peeling up from the bottom would involve projecting the original data downward to the lower boundary of the data, truncating off any lower left hand side horizontal region, running an OLS regression through the truncated projected data

to find a TPLS estimate for the observations that determined the lower boundary of the data, eliminating those observations that determined the lower boundary, and then reiterating this process until there are fewer than 10 observations left at the top of the data. By peeling the data from both the top to the bottom and from the bottom to the top, the observations at both the top and the bottom of the data will have an influence on the results. Of course, some of the observations in the middle of the data will have two TPLS estimated slopes associated with them—one from peeling the data downward and the other from peeling the data upward.

Above, we discussed projecting the data upward and downward; however, an alternative procedure would project the data to the left and to the right. 4D-RTPLS projects the data 4 different ways, upwards when peeling the data from the top, downward when peeling the data from the bottom, leftward when peeling the data from the left, and rightward when peeling the data from the right. When peeling the data from the right or left, any vertical sections of the frontier are truncated off for the same reasons that horizontal regions were truncated off when peeling the data downward and upward.

Once the entire dataset has been peeled from the top, bottom, left, and right, all the resulting TPLS estimates (with their associated data) are put into a final dataset. These TPLS estimates are then made the dependent variable in a final regression in which $1/X$ and $Y/X$ are the explanatory variables. The data are plugged back into this final regression to produce a separate 4D-RTPLS estimate for each observation. To understand the role of the final regression, consider Figure 1 again. If all the observations on the upper frontier had been associated with exactly the same omitted variable values (perhaps 98), then the resulting TPLS estimate would perfectly fit all of the observations it was associated with. However, Figure 1 shows that the observations on the upper frontier were associated with omitted variable values of 92, 98, 98, and 95. The resulting TPLS slope estimate would perfectly fit a $q$ value of approximately[5] 96 (the mean of 92, 98, 98, and 95). When a TPLS estimate for a $q$ of 96 is associated with qs of 92, 98, 98, and 95, some random variation (both positive and negative variation) remains. By combining the results from all iterations when peeling down, up, right, and left and then conducting this final regression, this random variation is eliminated.

Realize that $Y$ is codetermined by $X$ and $q$. Thus the combination of $X$ and $Y$ should contain information about $q$. This final regression exploits this insight in order to better capture the influence of $q$. The exact form of this final regression is justified by the following derivation.

In (2.2), the part usually omitted ($\alpha_2 X^n q^m$) could be of many different functional forms ("$n$" and "$m$" could be any real number, positive, or negative):

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 X^n q^m, \tag{2.2}$$

$$\frac{\partial Y}{\partial X} = \alpha_1 + n\alpha_2 X^{n-1} q^m \quad \text{(derivative of (2.2))}, \tag{2.3}$$

$$\frac{Y}{X} = \frac{\alpha_0}{X} + \alpha_1 + \alpha_2 X^{n-1} q^m \quad \text{(dividing (2.2) by } X\text{)}, \tag{2.4}$$

$$\alpha_1 + \alpha_2 X^{n-1} q^m = \frac{Y}{X} - \frac{\alpha_0}{X} \quad \text{(rearranging (2.4))}, \tag{2.5}$$

$$\frac{\partial Y}{\partial X} = fn\left(\frac{Y}{X}, \frac{1}{X}\right) \quad \text{(from (2.3) and (2.5))}. \tag{2.6}$$

If $n = 1$, then the right hand side of (2.3) perfectly matches the left hand side of (2.5) implying that just $Y/X$ and $1/X$ should be in (2.6). However, if $n \neq 1$, including either $Y$ or $Y$ and $X$ might produce better estimates.[6]

The mathematical equations used to calculate the frontier for each iteration of 4D-RTPLS are as follows: denote the dependent variable of observation "$i$" by $Y_i$, $i = 1, \ldots, I$, and the known independent variable of that observation by $X_i$, $i = 1, \ldots, I$. Consider the following variable returns to scale, output-oriented DEA problem, which is used when peeling the data downward:

$$\begin{aligned} &\max \Phi \\ \text{subject to} \quad &\Sigma_i \lambda_i X_i \leq X^o \\ &\Phi Y^o \leq \Sigma_i \lambda_i Y_i \\ &\Sigma_i \lambda_i = 1; \ \lambda_i \geq 0, \ i = 1, \ldots, I. \end{aligned} \tag{2.7}$$

The ratio of maximally expanded dependent variable to the actual dependent variable ($\Phi$) provides a measure of the influence of unfavorable omitted variables on each observation. This problem is solved $I$ times, once for each observation in the sample. For observation "$\circ$" under evaluation, the problem seeks the maximum expansion of the dependent variable $Y^\circ$ consistent with best practice observed in the sample, that is, subject to the constraints in the problem. In order to project each observation upward to the frontier, its $Y$ value is multiplied by $\Phi$ (for (2.7), $\Phi$ will be greater than or equal to 1). Peeling the data from the right is accomplished by using (2.7) after switching the positions of $X$ and $Y$ (in other words, every $X$ in (2.7) would refer to the dependent variable and every $Y$ in (2.7) would refer to the independent variable when peeling from the right side).

The variable returns to scale, input-oriented DEA problem used when peeling the data from the left is

$$\begin{aligned} &\min \Phi \\ \text{subject to} \quad &\Sigma_k \lambda_k Y_k \geq Y_i \\ &\Phi X_i \geq \Sigma_k \lambda_k X_k \\ &\Sigma_i \lambda_i = 1; \quad \lambda_k \geq 0, \quad k = 1, \ldots, I. \end{aligned} \tag{2.8}$$

To project the data to the frontier when peeling from the left, the $X$ value for each observation should be multiplied by $\Phi$ (for (2.8), $\Phi$ will be less than or equal to 1). Observations on the frontier will have a $\Phi = 1$ for both (2.7) and (2.8). Finally, to peel the data upward from the bottom, (2.8) will be used after switching the positions of $Y$ and $X$.

4D-RTPLS projected the data up, down, left, and right. However, if a plot of the data shows a tall and thin column, then it might be best to just project up and down. For example, if $q$ has a relatively large effect on the true slope, then the data will appear as a tall column with more efficient observations at the top of this column than at the sides. By projecting the data up and down, the data will be projected to where the efficient points are more concentrated. The more concentrated the efficient points are, the more likely they are to have similar $q$ values and thus the resulting TPLS estimates will be more accurate. In this case, UD-RTPLS

(Up Down RTPLS which only projects up and down) will produce better estimates than 4D-RTPLS, ceteris paribus.

For similar reasons, when $q$ has a relatively small effect on the true slope, the data will appear flat and fat, the efficient points will tend to be concentrated on the sides of the data, and LR-RTPLS (Left Right RTPLS) is likely to produce better estimates than 4D-RTPLS. Any round-off and measurement error that adds vertically to the value of $Y$ would decrease the accuracy of UD-RTPLS more than it decreased the accuracy of LR-RTPLS (because LR-RTPLS would not be going the same direction as the error was added). BP-RTPLS (best projection RTPLS) merely picks the direction of projection (UD, LR, or 4D) that produces the best estimates.

BP-RTPLS generates reduced form estimates that include all the ways that $X$ and $Y$ are correlated. Thus, even when many variables interact via a system of equations, a researcher using BP-RTPLS does not have to discover and justify that system of equations. In contrast, traditional regression analysis theoretically must include all relevant variables in the estimation and the resulting slope estimate for $dy/dx$ is for the effects of just $x$-holding all other variables constant. BP-RTPLS' reduced form estimates are not substitutes for traditional regression analysis' partial derivative estimates. Instead BP-RTPLS and traditional regression estimates are compliments which capture different types of information. BP-RTPLS has the disadvantage of not being able to tell the researcher the mechanism by which $X$ affects $Y$. On the other hand, BP-RTPLS has the advantage of not having to model and find data for all the forces that can affect $Y$ in order to estimate $\partial Y/\partial X$. Both BP-RTPLS and traditional regression techniques find "correlations." It is impossible for either one of them to prove "causation."

A brief survey of the literature leading up to BP-RTPLS is now provided.[7] Branson and Lovell [3] introduce the idea that by drawing a line around the top of a dataset and projecting the data to this line, one can eliminate variations in $Y$ that are due to variations in omitted variables. Branson and Lovell projected the data to the left, they did not truncate off any vertical section of the frontier, nor did they use a reiterative process. Leightner [4] projected the data upward, discovered that truncating off any horizontal section of the frontier improved the results, and instituted a reiterative process. He named the resulting procedure "Reiterative Truncated Projected Least Squares" (RTPLS).

Leightner and Inoue [5] ran simulation tests which show that RTPLS produces (on average) less than half the error of OLS when there are omitted variables that interact with the included variables under a wide range of conditions. Leightner and Inoue [5] also explain how situations where $Y$ is negatively related to $X$ can be handled, how omitted variables that can change the sign of the slope can be handled, and how the influence of additional right hand variables can be eliminated before conducting RTPLS. Leightner [6] introduces bidirectional reiterative truncated least squares (BD-RTPLS) which peeled the data from both the top and the bottom. Leightner [7] shows how the central limit theorem can be used to generate confidence intervals for groups of BD-RTPLS estimates. Published studies that used either RTPLS or BD-RTPLS in applications include Leightner [4, 6–12] and Leightner and Inoue [5, 13–15].

## 3. A Theoretical Argument That BP-RTPLS Is Unbiased

We will begin this section by explaining the conditions under which BP-RTPLS produces estimates that perfectly equal the true value of the slope. We will then argue that relaxing those conditions does not introduce bias into BP-RTPLS estimates. Therefore we will conclude that BP-RTPLS produces unbiased estimates. Figure 2 will be used to illustrate our argument.

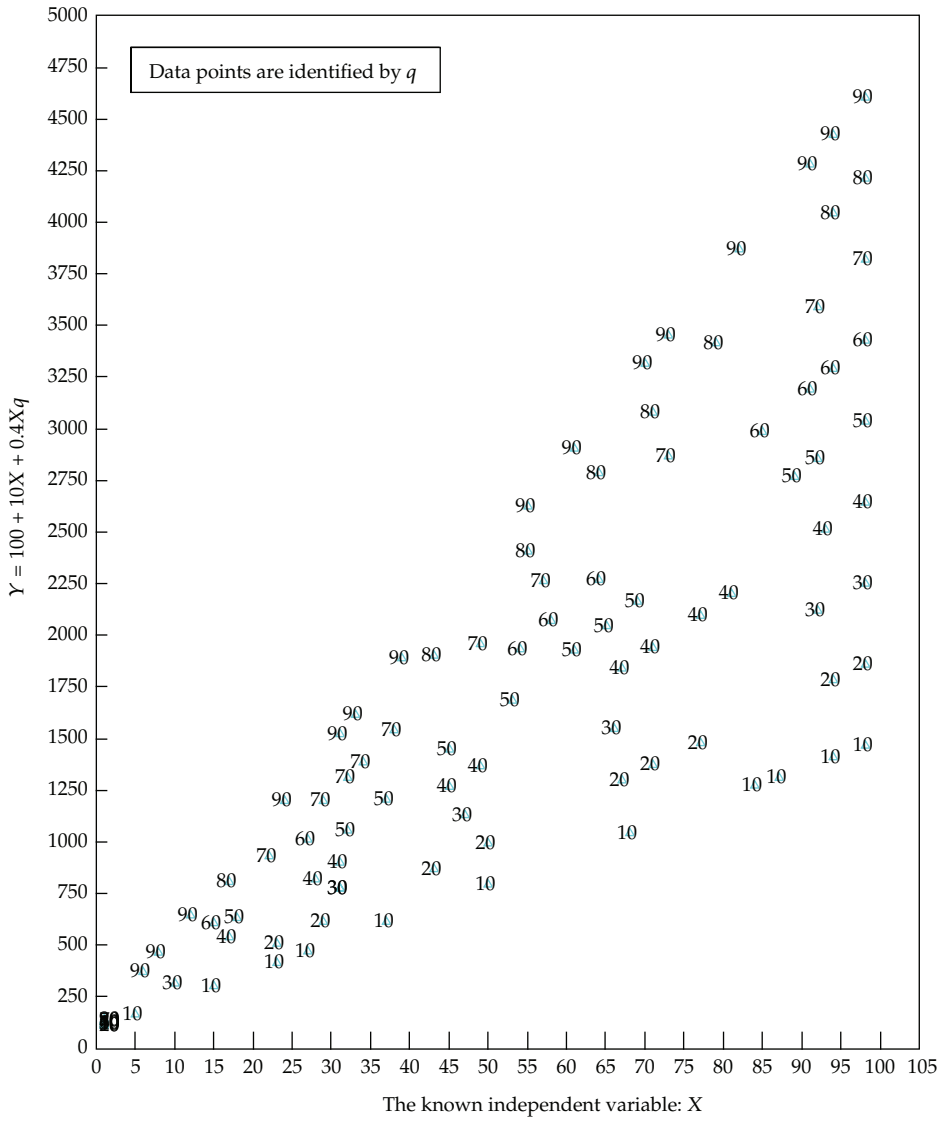



**Figure 2:** When TPLS works perfectly.

If there is no measurement and round-off error and if the smallest value and largest values for the known independent variable are associated with every possible value for the omitted variable, $q$, then UD-RTPLS, LR-RTPLS, 4D-RTPLS, and BP-RTPLS will all produce the same estimates which perfectly match the true slope. Figure 2 was generated by making $qs$ a member of the set $\{90, 80, 70, \ldots, 10\}$, associating the smallest $X$, which had the value of 1, with each of those $qs$ and then associating the largest $X$, which had the value of 98, with each of those $qs$. The remaining observations were created by randomly generating $Xs$ between 1 and 98 and randomly associating one of the $qs$ with each observation.

In Figure 2, the first iteration when peeling the data downward would produce the true slope for all of the observations that determined the frontier in that iteration. For both Figures 1 and 2, $Y = 100 + 10X + 0.4qX$; thus $\partial Y/\partial X = 10 + 0.4q = 10 + 0.4(90) = 46$ for

the first iteration. The second iteration will also find the true slope for the observations on its frontier—a slope of 10 + 0.4(80) = 42. This will be true for all iterations. Furthermore, the exact same perfect slope will be found when the data are projected to the left when peeling from the left. Moreover, when peeling the data upwards and from the right, all iterations will continue to produce a perfect slope. The reason that each iteration works perfectly is that the two ends of each frontier contain identical omitted variable values which correspond to the largest (when peeling down or from the left) or smallest (when peeling up or from the right) omitted variable values remaining in the dataset; thus a frontier between the smallest and largest $Xs$ will be a straight line with a slope that perfectly matches the true $\partial Y/\partial X$ of every observation on the frontier. In this case, there is no need to run the final regression of BP-RTPLS because each TPLS estimate is perfect. However if that final regression is run any way, it will produce a $R$-squared of 1.0 and plugging the data back into the resulting equation will regenerate the TPLS estimate from each iteration.

Now that we have established under what conditions BP-RTPLS produces estimates that perfectly match the true slope, we will discuss what happens when those conditions are not met. Changes in these conditions can be grouped into three categories: (1) changes for which the TPLS estimates continue to perfectly match the true slope, (2) changes that will produce TPLS estimates that are greater than the true slope for observations with relatively small $Xs$ and that are less than the true slope for observations with relatively large $Xs$, and (3) changes for which all the TPLS estimates of a given iteration are greater than (or less than) the true slope. We will provide reasons why each of these types of changes will not introduce systematic bias into the final BP-RTPLS estimates.

Omitting an observation from the middle of the frontier will not affect the TPLS slope estimates (to see this, eliminate any, or all, of the middle of the frontier observations that correspond to a $q$ of 90 in Figure 2). Likewise, if the observation corresponding to the upper right hand 90 in Figure 2 is eliminated, then the first iteration when peeling the data downward would continue to generate the true slope because eliminating that observation would just create a small horizontal region in the first iteration which would be truncated off.

However, if the three observations for $q = 90$ in the upper right part of Figure 2 were all eliminated, then the observation identified by an 80 in the upper right would define the upper right side of the first frontier. In this case, the resulting TPLS estimate of the slope for the first iteration would be slightly too small for the observations identified with 90 s and too big for the upper most observation identified by an 80.[8] The same phenomenon happens when we are peeling upward (or from the right), if the observation identified by a $q$ of 10 on the right hand side was eliminated. In this case the observation identified by a 20 on the far right side would define the right side of the first frontier; as a consequence, the first iteration when peeling upward (or from the right) would generate a slope that was slightly too large for the observations with a $q$ of 10 but too small for the observation with a $q$ of 20. In both of these cases, the TPLS estimated slope of the observations with relatively small $Xs$ are too large and the TPLS estimated slope of the observations with relatively large $Xs$ are too small. It is important to note that, since the TPLS slope estimate for this iteration is found using OLS, the relative weight of the slopes overestimated in this iteration should approximately equal the relative weight of the slopes underestimated. The relative weight of the overestimation would cancel out with the relative weight of the underestimation when the final regression of the BP-RTPLS process forces the results to go through the origin, thus eliminating any possible bias from this phenomenon.[9]

The third type of changes in Figure 2 would cause all of the TPLS estimates for a given iteration to be larger than (or smaller than) the true slope. For example, when the

dataset is peeled downward (or from the left) if all the observations corresponding to $X = 1$ were eliminated, then the lower left hand observation identified by a 10 would define the lower left edge of the first frontier. In this case TPLS would generate a slope estimate that was slightly too large for the observations identified by 90 s and much too large for the one observation identified by the 10. Likewise when peeling the data upwards (or from the right) if all of the observations identified with an $X = 1$ were eliminated and the next two observations identified by a 10 (in the lower left part of Figure 2) were eliminated, then the observation identified by a 30 in the lower left side of Figure 2 would define the left hand edge of the first frontier. In this case the TPLS slope estimate would be slightly too small for all the observations identified by a 10 on the frontier and much too small for the observation identified by the 30. The incidence and weight of TPLS estimates that are greater than the true slope should be approximately equal to the incidence and weight of TPLS estimates that are less than the true slope when the final BP-RTPLS estimate is made. Thus these inaccuracies in the TPLS estimates should also be eliminated when the final BP-RTPLS estimate is made.

None of the three categories of changes discussed above would add a systematic bias to BP-RTPLS estimates. Additional types of changes are possible, like eliminating observations on both ends of the frontier for a given iteration; however, these types of changes would cause effects that are some combination of the effects discussed above. Finally there is no reason why "random" error would add systematic bias either.

## 4. Simulation Results

Our first set of simulations are based on computer generated values of $X$ and $q$ which are uniform random numbers $\sim U[0, 10]$, where 0 is the lower bound of the distribution and 10 is the upper bound. Measurement and round-off error, $e$, is generated as a normal random number whose standard deviation is adjusted to be 0%, 1%, or 10% of variable $X$'s standard deviation. We consider 18 cases—all the combinations where (1) the omitted variable ($q$) makes a 10%, 100%, or a 1000% difference in $\partial Y/\partial X$, (2) where measurement and round-off error is 0%, 1%, or 10% of $X$, and (3) either 100 observations or 500 observations are used. Equations (4.1), (4.2), and (4.3) are used to model when the omitted variable makes a 10%, 100%, and 1000% difference in $\partial Y/\partial X$, respectively.

Consider

$$Y = 10 + 1.0X + 0.01qX + e, \tag{4.1}$$

$$Y = 10 + 1.0X + 0.1qX + e, \tag{4.2}$$

$$Y = 10 + 1.0X + 1.0qX + e, \tag{4.3}$$

$\partial Y/\partial X$ for (4.2) would be $1 + 0.1q$; since $q$ ranges from 0 to 10, the true slope will range from 1 (when $q = 0$) to 2 (when $q = 10$). Thus, for (4.2), the omitted variable, $q$, makes a 100% difference to the true slope. For similar reasons $q$ makes a 10% difference to the real slope in (4.1) and approximately a 1000% difference in (4.3). Total error for the $i$th observation would equal the error from the omitted variable plus the added measurement and round-off error.

Tables 1 and 2 present the mean of the absolute value of the error and the standard deviation of the error for 18 sets of 5000 simulations each where the errors from OLS and from [][]RTPLs are defined by (4.4) and (4.5), respectively. In these equations, "OLS" refers to the

OLS estimate of $\partial Y/\partial X$ when $q$ is omitted and "True" refers to the true slope as calculated by plugging each observation's data into the derivatives of (4.1)–(4.3) above. "[][]RTPLS" is the [][]RTPLS estimate of $\partial Y/\partial X$, where "BD," "UD," "LR," or "4D" could be substituted for "[][]."

$$\text{Define } E_i^{\text{ols}} = \frac{(\text{OLS} - \text{True}_i)}{\text{True}_i}. \tag{4.4}$$

$$\text{Define } E_i^{[\,][\,]\text{RTPLS}} = \frac{([\,][\,]\text{RTPLS} - \text{True}_i)}{\text{True}_i}. \tag{4.5}$$

The mean absolute value of the percent OLS error (Table 1, row 1) was calculated from (4.6), where "$n$" is the number of observations in a simulation and "$m$" is the number of simulations:

$$\frac{\sum_{j=1}^{m}\left[\sum_{i=1}^{n}\left|E_i^{\text{OLS}}\right|/n\right]}{m}. \tag{4.6}$$

Equation (4.7) was used to calculate the standard deviation of OLS error (Table 2, row 1), where $E(E_i^{\text{OLS}})$ = the mean of $E_i^{\text{OLS}} = (\sum_{i=1}^{n} E_i^{\text{OLS}})/n$.

Consider

$$\frac{\left\{\sum_{j=1}^{M}\left[\left(\sum_{i=1}^{n}\left(E_i^{\text{OLS}} - E\left(E_i^{\text{OLS}}\right)\right)^2\right)/(n-1)\right]^{1/2}\right\}}{m}. \tag{4.7}$$

The absolute value of the mean error (Table 1) and the standard deviation (Table 2) of [][]RTPLS error (Row 2) were calculated with (4.5)–(4.7), respectively, where "$E_i^{[][]\text{RTPLS}}$" was substituted for "$E_i^{\text{OLS}}$."

The results when 100 observations are used in each simulation are shown in Panel A, and the results when 500 observations are used are shown in Panel B. Columns 1–3, 4–6, and 7–9 correspond to when the omitted variable makes a 10%, 100%, and 1000% difference in $\partial Y/\partial X$, respectively. No measurement and round-off error was added for columns 1, 4, and 7; 1% measurement and round-off error was added for columns 2, 5, and 8; and 10% measurement and round-off error was added for columns 3, 6, and 9. Row one of Tables 1 and 2 presents the OLS results when $q$ was omitted. Row 2a presents the results of using BD-RTPLS, the second generation of this technique.[10] Rows 2b, 2c, and 2d present the results of using UD-RTPLS, LR-RTPLS, and 4D-RTPLS, respectively. When running the simulations for rows 2b, 2c, and 2d, three different sets of possible explanatory variables for the final regression were considered: $\{1/X, Y/X\}$, $\{1/X, Y/X, Y\}$, and $\{1/X, Y/X, Y, X\}$. The set of final regression explanatory variables that produced the largest OLS/[][]RTPLS ratio for rows 2b, 2c, and 2d of a given column is what is reported in that column for Tables 1 and 2. This set of final regression explanatory variables was $1/X$, $Y/X$, $Y$, and $X$ for column 3 and just $1/X$ and $Y/X$ for all other columns. Row 2e and 3e for BP-RTPLS (Best Projection-RTPLS) just repeats the result in the three lines above it that corresponds to the largest OLS/[][]RTPLS ratio.

**Table 1:** The mean of the absolute value of the error.

| Column number | 1 | 2 | 3[1] | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Importance of omitted $q$ | 10% | 10% | 10% | 100% | 100% | 100% | 1000% | 1000% | 1000% |
| Size of measurement $e$ | 0% | 1% | 10% | 0% | 1% | 10% | 0% | 1% | 10% |
| Panel A: 100 observations in each simulation; 5000 simulations | | | | | | | | | |
| (1) Mean % OLS error | 0.0240 | 0.0240 | 0.0249 | 0.1779 | 0.1779 | 0.1779 | 0.7143 | 0.7143 | 0.7143 |
| (2) Mean % [][]RTPLS error | | | | | | | | | |
| (a) BD-RTPLS*[2] | 0.0131 | 0.0191 | 0.0253 | 0.0787 | 0.0827 | 0.1347 | 0.4463 | 0.4469 | 0.4603 |
| (b) UD-RTPLS[3] | 0.0044 | 0.0182 | 0.0463 | 0.0311 | 0.0359 | 0.1557 | 0.0715 | 0.0717 | 0.0893 |
| (c) LR-RTPLS[4] | 0.0042 | 0.0179 | 0.0464 | 0.0424 | 0.0455 | 0.1441 | 0.1856 | 0.1858 | 0.1906 |
| (d) 4D-RTPLS[5] | 0.0042 | 0.0180 | 0.0462 | 0.0292 | 0.0335 | 0.1467 | 0.1119 | 0.1122 | 0.1225 |
| (e) BP-RTPLS[6] | 0.0042 | 0.0179 | | 0.0292 | 0.0335 | 0.1441 | 0.0715 | 0.0717 | 0.0893 |
| (3) OLS/[][]RTPLS error | | | | | | | | | |
| (a) BD-RTPLS* | 1.90 | 1.31 | 1.00 | 2.42 | 2.31 | 1.41 | 2.03 | 2.03 | 1.96 |
| (b) UD-RTPLS | 12.77 | 2.74 | 0.59 | 12.58 | 8.64 | 2.19 | **18.92** | **17.66** | **11.51** |
| (c) LR-RTPLS | **14.25** | **2.82** | 0.59 | 9.50 | 6.76 | **2.58** | 3.21 | 3.09 | 3.32 |
| (d) 4D-RTPLS | 12.97 | 2.80 | 0.59 | **13.78** | **9.34** | 2.54 | 11.03 | 9.93 | 7.83 |
| (e) BP-RTPLS | 14.25 | 2.82 | | 13.78 | 9.34 | 2.58 | 18.92 | 17.66 | 11.51 |
| Panel B: 500 observations in each simulation; 5000 simulations | | | | | | | | | |
| (1) Mean % OLS error | 0.0239 | 0.0239 | 0.0241 | 0.1769 | 0.1769 | 0.1769 | 0.7111 | 0.7111 | 0.7111 |
| (2) Mean % [][]RTPLS error | | | | | | | | | |
| (a) BD-RTPLS* | 0.0106 | 0.0215 | 0.0244 | 0.0563 | 0.0709 | 0.1535 | 0.3324 | 0.3336 | 0.3840 |
| (b) UD-RTPLS | 0.0026 | 0.0302 | 0.0472 | 0.0291 | 0.0426 | 0.2994 | 0.0374 | 0.0390 | 0.0885 |
| (c) LR-RTPLS | 0.0020 | 0.0289 | 0.0472 | 0.0483 | 0.0521 | 0.2574 | 0.1942 | 0.1944 | 0.1978 |
| (d) 4D-RTPLS | 0.0019 | 0.0295 | 0.0470 | 0.0144 | 0.0253 | 0.2759 | 0.0869 | 0.0873 | 0.1056 |
| (e) BP-RTPLS | 0.0019 | 0.0289 | | 0.0144 | 0.0253 | 0.2574 | 0.0374 | 0.0390 | 0.0885 |
| (3) OLS/[][]RTPLS error | | | | | | | | | |
| (a) BD-RTPLS* | 2.30 | 1.13 | 0.99 | 3.09 | 2.60 | 1.16 | 2.47 | 2.46 | 2.17 |
| (b) UD-RTPLS | 21.14 | 2.29 | 0.56 | 9.50 | 5.48 | 0.86 | **29.91** | **24.49** | **11.26** |
| (c) LR-RTPLS | 34.58 | **2.55** | 0.56 | 3.03 | 3.75 | **1.67** | 1.95 | 1.95 | 2.62 |
| (d) 4D-RTPLS | **39.79** | 2.45 | 0.57 | **34.86** | **14.74** | 1.27 | 5.25 | 5.31 | 7.37 |
| (e) BP-RTPLS | 39.79 | 2.55 | | 34.86 | 14.74 | 1.67 | 29.91 | 24.49 | 11.26 |

[1]Each row was calculated with the following three sets of explanatory variables for the final regression: $\{1/X, Y/X\}$, $\{1/X, Y/X, Y\}$, and $\{1/X, Y/X, Y, X\}$. Column 3 shows the results when $1/X$, $Y/X$, $Y$, and $X$ are used as the explanatory variables in the final regression because the approach with the greatest OLS/[][]RTPLS error ratio always used those variables for column 3. For all other columns, the approach with the greatest OLS/[][]RTPLS error ratio always used solely $1/X$ and $Y/X$ and the corresponding results are those reported here.

[2]BD-RTPLS*: UD-RTPLS except a constant is used in the final regression. Unlike BD-RTPLS, BD-RTPLS* does not truncate off the 3% of the observations corresponding to the smallest (largest) $X$s when peeling down (up).

[3]UD-RTPLS: RTPLS where the data are solely projected up and down, not left and right.

[4]LR-RTPLS: RTPLS where the data are solely projected to the left and right, not up and down.

[5]4D-RTPLS: RTPLS where the data are projected up, down, left, and right.

[6]BP-RTPLS: the results for the approach—UD-RTPLS, LR-RTPLS, or 4D-RTPLS—that produces the greatest OLS/[][]RTPLS ratio.

When comparing the relative absolute value of the mean error (Table 1) and standard deviation (Table 2) of OLS error to [][]RTPLS error by observation, "$\text{Ln}(|E_i^{\text{OLS}}|/|E_i^{[][]\text{RTPLS}}|)$" was substituted for $|E_i^{\text{OLS}}|$ in (4.6) and for $E_i^{\text{OLS}}$ in (4.7) and then the antilog of the result was found (row 3 of Tables 1 and 2, resp.).[11] The natural log of the ratio of OLS to [][]RTPLS error had to be used in order to center this ratio symmetrically around the number 1. Consider a two observation example where the ratio is 5/1 for one observation and 1/5 for the other

**Table 2:** The standard deviation of the error.

| Column number | 1 | 2 | 3[1] | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Importance of omitted $q$ | 10% | 10% | 10% | 100% | 100% | 100% | 1000% | 1000% | 1000% |
| Size of measurement $e$ | 0% | 1% | 10% | 0% | 1% | 10% | 0% | 1% | 10% |
| Panel A: 100 observations in each simulation; 5000 simulations | | | | | | | | | |
| (1) Percent OLS error | 0.0275 | 0.0275 | 0.0275 | 0.2097 | 0.2097 | 0.2097 | 1.0917 | 1.0917 | 1.0917 |
| (2) Percent [][]RTPLS error | | | | | | | | | |
| (a) BD-RTPLS*[2] | 0.0147 | 0.0220 | 0.0276 | 0.0906 | 0.0963 | 0.1621 | 0.6595 | 0.6605 | 0.6830 |
| (b) UD-RTPLS[3] | 0.0025 | 0.0391 | 0.0567 | 0.0226 | 0.0406 | 0.2249 | 0.1272 | 0.1297 | 0.1913 |
| (c) LR-RTPLS[4] | 0.0025 | 0.0392 | 0.0571 | 0.0240 | 0.0429 | 0.2369 | 0.1621 | 0.1649 | 0.2332 |
| (d) 4D-RTPLS[5] | 0.0025 | 0.0392 | 0.0569 | 0.0233 | 0.0418 | 0.2309 | 0.1445 | 0.1471 | 0.2121 |
| (e) BP-RTPLS[6] | 0.0025 | 0.0392 | 0.0571 | 0.0233 | 0.0418 | 0.2369 | 0.1272 | 0.1297 | 0.1913 |
| (3) OLS/[][]RTPLS | | | | | | | | | |
| (a) BD-RTPLS* | 0.67 | 0.70 | 0.72 | 0.75 | 0.77 | 0.75 | 1.09 | 1.09 | 1.08 |
| (b) UD-RTPLS | 1.11 | 1.54 | 1.48 | 1.16 | 1.27 | 1.43 | **1.40** | **1.41** | **1.49** |
| (c) LR-RTPLS | **1.11** | **1.55** | **1.48** | 1.12 | 1.22 | **1.51** | 1.24 | 1.24 | 1.31 |
| (d) 4D-RTPLS | 1.11 | 1.55 | 1.48 | **1.17** | **1.29** | 1.49 | 1.30 | 1.31 | 1.40 |
| (e) BP-RTPLS | 1.11 | 1.55 | 1.48 | 1.17 | 1.29 | 1.51 | 1.40 | 1.41 | 1.49 |
| Panel B: 500 observations in each simulation; 5000 simulations | | | | | | | | | |
| (1) Percent OLS error | 0.0275 | 0.0275 | 0.0275 | 0.2097 | 0.2097 | 0.2097 | 1.0944 | 1.0944 | 1.0944 |
| (2) Percent [][]RTPLS error | | | | | | | | | |
| (a) BD-RTPLS* | 0.0120 | 0.0251 | 0.0274 | 0.0645 | 0.0868 | 0.1857 | 0.5044 | 0.5066 | 0.5921 |
| (b) UD-RTPLS | 0.0011 | 0.0837 | 0.0581 | 0.0115 | 0.0645 | 0.3639 | 0.0610 | 0.0725 | 0.2445 |
| (c) LR-RTPLS | 0.0011 | 0.0840 | 0.0589 | 0.0120 | 0.0695 | 0.3942 | 0.0852 | 0.0987 | 0.3037 |
| (d) 4D-RTPLS | 0.0011 | 0.0839 | 0.0585 | 0.0117 | 0.0670 | 0.3790 | 0.0730 | 0.0855 | 0.2740 |
| (e) BP-RTPLS | 0.0011 | 0.0840 | 0.0589 | 0.0117 | 0.0670 | 0.3942 | 0.0610 | 0.0725 | 0.2445 |
| (3) OLS/[][]RTPLS error | | | | | | | | | |
| (a) BD-RTPLS* | 0.64 | 0.56 | 0.75 | 0.87 | 0.85 | 0.62 | 0.88 | 0.88 | 0.87 |
| (b) UD-RTPLS | 1.05 | 1.49 | 1.48 | 1.07 | 1.18 | 1.21 | **1.38** | **1.40** | **1.48** |
| (c) LR-RTPLS | 1.07 | **1.53** | **1.48** | 1.04 | 1.14 | **1.36** | 1.25 | 1.25 | 1.32 |
| (d) 4D-RTPLS | **1.07** | 1.52 | 1.48 | **1.11** | **1.38** | 1.29 | 1.24 | 1.25 | 1.41 |
| (e) BP-RTPLS | 1.07 | 1.53 | 1.48 | 1.11 | 1.38 | 1.36 | 1.38 | 1.40 | 1.48 |

[1]Each row was calculated with the following three sets of explanatory variables for the final regression: $\{1/X, Y/X\}$, $\{1/X, Y/X, Y\}$, and $\{1/X, Y/X, Y, X\}$. Column 3 shows the results when $1/X$, $Y/X$, $Y$, and $X$ are used as the explanatory variables in the final regression because the approach with the greatest OLS/[][]RTPLS error ratio always used those variables for column 3. For all other columns, the approach with the greatest OLS/[][]RTPLS error ratio always used solely $1/X$ and $Y/X$ and the corresponding results are those reported here.

[2]BD-RTPLS*: UD-RTPLS except a constant is used in the final regression. Unlike BD-RTPLS, BD-RTPLS* does not truncate off the 3% of the observations corresponding to the smallest (largest) $X$s when peeling down (up).

[3]UD-RTPLS: RTPLS where the data are solely projected up and down, not left and right.

[4]LR-RTPLS: RTPLS where the data are solely projected to the left and right, not up and down.

[5]4D-RTPLS: RTPLS where the data are projected up, down, left, and right.

[6]BP-RTPLS: the results for the approach—UD-RTPLS, LR-RTPLS, or 4D-RTPLS—that produces the greatest OLS/[][]RTPLS ratio.

observation. In this example, the mean OLS/[][]RTPLS ratio is 2.6 making OLS appear to have 2.6 times as much error as [][]RTPLS, when (in this example) OLS and [][]RTPLS are performing the same on average. Taking the natural log solves this problem. Ln(5) = 1.609 and Ln(1/5) = −1.609 and their average would be zero and the antilog of zero is 1, correctly showing that OLS and [][]RTPLS are performing equally well in this example.

In our tables, we present the mean of the absolute value of the error for OLS and for [][]RTPLS so that the reader can understand the size of the error involved. However, our primary focus is on the OLS/[][]RTPLS ratio because this ratio gives the greatest possible emphasis on the accuracy of estimates for individual observations. It is important to realize that dividing the mean absolute value of the error for OLS by the mean absolute value of the error for [][]RTPLS will not duplicate the OLS/[][]RTPLS error ratio.

Table 1 shows that the mean of the absolute value of the error from OLS is 2.4% to 2.5% when $q$ makes a 10% difference to the true slope (Panel A, line 1, columns 1–3); in contrast, when $q$ makes a 1000% difference to the true slope, the mean error from OLS is 71.4% (Panel A, line 1, columns 7-8). In contrast, the mean of the absolute value of the error from BD-RTPLS is only 8.93% when $q$ makes a 1000% difference and $e$ = 10% (Panel A, line 2b, column 9). Moving from 71.4% error to 8.9% error is a huge improvement.

Notice also that the mean of the absolute value of error for OLS does not noticeably change with the amount of measurement and round-off error added, but the mean of [][]RTPLS error does increase as measurement and round-off error increases (Table 1, lines 1 and 2). Furthermore, as the sample size increases from 100 observations (Panel A) to 500 observations (Panel B), the mean of the absolute value of OLS error does not noticeably fall; however, sometimes the mean [][]RTPLS error falls and sometimes it rises as the sample size increases from 100 to 500 observations. We have no convincing explanation for why the mean [][]RTPLS error sometimes rises as the sample size increases.

OLS produces greater mean error than [][]RTPLS except for when $q$ = 10% and $e$ = 10% for both sample sizes (lines 1 and 2, column 3) and when $q$ = 10%, $e$ = 1%, and when $q$ = 100%, $e$ = 10% when 500 observations are used (lines 1-2, columns 2 and 6, Panel B). When we focus on the OLS/[][]RTPLS mean error ratio, [][]RTPLS outperforms OLS for all cases (the OLS/[][]RTPLS ratio is greater than 1) except for when $q$ only makes a 10% difference and $e$ = 10%. It makes sense that when $q$ and $e$ are the same size, then [][]RTPLS is not able to use the relative vertical position of observations to capture the influence of $q$ (because this vertical position contains an equal amount of $e$ contamination).

When 100 observations and the best projection direction is used (line 2$e$), the OLS/[][]RTPLS ratio shows (ignoring the case where both $q$ and $e$ = 10%) that OLS produces between 2.58 times to 18.92 times (258% to 1892%) more error than [][]RTPLS. When 500 observations and the best projection direction are used, (ignoring the case where both $q$ and $e$ = 10%), OLS produces between 1.67 times to 39.79 times (167% to 3979%) more error than [][]RTPLS.

Table 1 (line 3) reveals a very interesting pattern. The optimal projection direction is left and right (LR-RTPLS) when $q$ makes a 10% difference and $e$ = 1%; is left, right, up, and down (4D-RTPLS) when $q$ makes a 100% difference and $e$ = 0% or 1%; is again left and right when $q$ = 100% and $e$ = 10%; and is always up and down (UD-RTPLS) when $q$ makes a 1000% difference. This pattern is the same for 100 observations and 500 observations and is the exact same pattern that is obtained by looking at the maximum OLS/[][]RTPLS ratios for the standard deviation of the error (Table 2, line 3). Furthermore, this pattern reappears in Tables 3 and 4 (Panel B) when a single set of data is extensively analyzed. This is a persistent pattern.

As discussed in Section 2 of this paper, an increase in the importance of $q$ should stretch the data upwards, leading to the efficient observations being more concentrated at the top of the frontier than they are along the sides of the frontier, which would cause a projection upward and downward (UD-RTPLS) to be more accurate than a projection left or right—concentrated efficient observations must have more similar values for $q$ than nonconcentrated

**Table 3:** One set of data, $Y = 5 + X + \alpha Xq + 0.4e$.

| | | | Mean error | | | | Mean OLS/[][]RTPLS $e$ | | |
|---|---|---|---|---|---|---|---|---|---|
| Row | $q$% | $e$% of $Y$ | OLS | UD | LR | 4D | UD | LR | 4D |
| Panel A (1/$X$, $Y/X$, $Y$, and $X$ in final regression) | | | | | | | | | |
| 1 | 120% | 15.89% | 0.1893 | 0.1668 | 0.1797 | 0.1704 | 0.96 | 0.92 | 0.97 |
| 2 | 130% | 15.35% | 0.2003 | 0.1719 | 0.1828 | 0.1748 | **1.10** | 1.01 | 1.06 |
| 3 | 140% | 14.84% | 0.2108 | 0.1805 | 0.1871 | 0.1805 | 1.05 | 1.06 | **1.13** |
| 4 | 150% | 14.37% | 0.2209 | 0.1909 | 0.1976 | 0.1910 | 1.06 | 1.15 | **1.19** |
| 5 | 160% | 13.92% | 0.2307 | 0.1927 | 0.1964 | 0.1913 | 1.03 | 1.15 | **1.15** |
| 6 | 170% | 13.50% | 0.2401 | 0.1987 | 0.2016 | 0.1968 | 1.07 | 1.17 | **1.18** |
| 7 | 180% | 13.11% | 0.2492 | 0.2028 | 0.2034 | 0.1999 | 1.11 | 1.22 | **1.28** |
| 8 | 190% | 12.74% | 0.2580 | 0.2032 | 0.2086 | 0.2020 | 1.12 | 1.18 | **1.28** |
| 9 | 200% | 12.39% | 0.2666 | 0.2112 | 0.2155 | 0.2090 | 1.13 | 1.23 | **1.30** |
| 10 | 210% | 12.06% | 0.2748 | 0.2174 | 0.2195 | 0.2144 | 1.12 | **1.29** | 1.27 |
| 11 | 220% | 11.74% | 0.2829 | 0.2237 | 0.2230 | 0.2192 | 1.16 | **1.35** | 1.31 |
| 12 | 230% | 11.44% | 0.2906 | 0.2293 | 0.2289 | 0.2245 | 1.20 | 1.26 | **1.30** |
| 13 | 240% | 11.16% | 0.2982 | 0.2300 | 0.2258 | 0.2236 | 1.17 | **1.32** | 1.28 |
| 14 | 250% | 10.89% | 0.3056 | 0.2348 | 0.2289 | 0.2277 | 1.23 | **1.36** | 1.30 |
| 15 | 260% | 10.63% | 0.3127 | 0.2361 | 0.2319 | 0.2293 | 1.19 | 1.27 | **1.32** |
| 16 | 270% | 10.38% | 0.3197 | 0.2399 | 0.2311 | 0.2313 | 1.19 | 1.29 | **1.32** |
| 17 | 280% | 10.15% | 0.3265 | 0.2438 | 0.2331 | 0.2343 | 1.22 | **1.35** | 1.32 |
| 18 | 290% | 9.93% | 0.3331 | 0.2480 | 0.2371 | 0.2377 | 1.20 | 1.31 | **1.38** |
| 19 | 300% | 9.71% | 0.3396 | 0.2534 | 0.2430 | 0.2444 | 1.24 | 1.31 | 1.30 |
| Panel B (1/$X$ and $Y/X$ in final regression) | | | | | | | | | |
| 20 | 270% | 10.38% | 0.3197 | 0.4286 | 0.3942 | 0.4113 | 0.65 | 0.84 | 0.73 |
| 21 | 280% | 10.15% | 0.3265 | 0.4061 | 0.3759 | 0.3909 | 0.74 | 1.01 | 0.84 |
| 22 | 290% | 9.93% | 0.3331 | 0.3913 | 0.3587 | 0.3741 | 0.81 | 1.16 | 0.99 |
| 23 | 300% | 9.71% | 0.3396 | 0.3740 | **0.3442** | 0.3577 | 0.93 | **1.32** | 1.18 |
| 24 | 320% | 9.31% | 0.3520 | 0.3651 | **0.3282** | 0.3447 | 0.97 | **1.49** | 1.29 |
| 25 | 340% | 8.94% | 0.3639 | 0.3539 | **0.3094** | 0.3284 | 1.04 | **1.83** | 1.40 |
| 26 | 360% | 8.60% | 0.3753 | 0.3133 | **0.2901** | 0.2989 | 1.58 | **2.82** | 2.04 |
| 27 | 380% | 8.28% | 0.3862 | 0.3030 | **0.2810** | 0.2862 | 1.70 | **2.58** | 2.48 |
| 28 | 390% | 8.13% | 0.3915 | 0.2907 | 0.2786 | **0.2775** | 1.92 | 2.46 | **2.97** |
| 29 | 400% | 7.99% | 0.3966 | 0.2945 | 0.2758 | **0.2744** | 1.85 | 2.49 | **2.77** |
| 30 | 420% | 7.71% | 0.4067 | 0.2812 | 0.2701 | **0.2654** | 2.02 | 2.53 | **2.99** |
| 31 | 440% | 7.46% | 0.4163 | 0.2613 | 0.2641 | **0.2586** | 2.87 | 2.66 | **3.06** |
| 32 | 450% | 7.34% | 0.4210 | **0.2556** | 0.2640 | 0.2564 | **3.44** | 2.51 | 2.84 |
| 33 | 460% | 7.22% | 0.4256 | **0.2518** | 0.2659 | 0.2547 | **3.32** | 2.43 | 2.78 |
| 34 | 480% | 6.99% | 0.4346 | **0.2476** | 0.2664 | 0.2496 | **3.39** | 2.44 | 3.05 |
| 35 | 500% | 6.78% | 0.4432 | **0.2404** | 0.2672 | 0.2496 | **3.13** | 2.16 | 2.71 |

**Table 4:** One set of data, additional simulations.

| | | | Mean error | | | | Mean OLS/[][]RTPLS error | | |
|---|---|---|---|---|---|---|---|---|---|
| Row | $q$% | $e$% of $Y$ | OLS | UD | LR | 4D | UD | LR | 4D |
| Panel A (1/$X$, $Y/X$, $Y$, and $X$ in final regression) $e$ = 20% of $X$ | | | | | | | | | |
| 1 | 40% | 10.94% | 0.0795 | 0.0892 | 0.0952 | 0.0918 | 0.80 | 0.74 | 0.77 |
| 2 | 50% | 10.43% | 0.0958 | 0.0940 | 0.1028 | 0.0977 | **1.12** | 0.85 | 0.92 |
| 3 | 60% | 9.97% | 0.1112 | 0.1002 | 0.1062 | 0.1019 | 1.01 | 0.95 | **1.04** |
| 4 | 70% | 9.55% | 0.1258 | 0.1073 | 0.1151 | 0.1093 | **1.20** | 1.07 | 1.19 |
| 5 | 80% | 9.16% | 0.1396 | 0.1155 | 0.1210 | 0.1164 | 1.14 | 1.12 | **1.21** |
| 6 | 90% | 8.80% | 0.1527 | 0.1248 | 0.1354 | 0.1279 | 1.16 | 1.20 | **1.21** |
| 7 | 100% | 8.47% | 0.1652 | 0.1339 | 0.1430 | 0.1369 | **1.25** | 1.25 | 1.23 |
| 8 | 110% | 8.16% | 0.1771 | 0.1411 | 0.1463 | 0.1417 | 1.28 | 1.25 | **1.44** |
| 9 | 120% | 7.88% | 0.1885 | 0.1501 | 0.1547 | 0.1500 | 1.35 | 1.25 | 1.42 |
| Panel B (1/$X$ and $Y/X$ in final regression) $e$ = 20% of $X$ | | | | | | | | | |
| 10 | 100% | 8.47% | 0.1652 | 0.2901 | 0.2641 | 0.2762 | 0.69 | 0.97 | 0.82 |
| 11 | 110% | 8.16% | 0.1771 | 0.2732 | 0.2511 | 0.2613 | 0.88 | 1.17 | 1.02 |
| 12 | 120% | 7.88% | 0.1885 | 0.2632 | **0.2381** | 0.2495 | 0.98 | **1.56** | 1.17 |
| 13 | 130% | 7.61% | 0.1995 | 0.2490 | **0.2267** | 0.2357 | 1.18 | **1.78** | 1.49 |
| 14 | 140% | 7.36% | 0.2100 | 0.2318 | **0.2186** | 0.2230 | 1.46 | **1.95** | 1.88 |
| 15 | 150% | 7.13% | 0.2202 | 0.2196 | 0.2129 | **0.2130** | 2.00 | 2.06 | **2.15** |
| 16 | 160% | 6.91% | 0.2299 | 0.2124 | 0.2097 | **0.2071** | 2.08 | 1.96 | **2.25** |
| 17 | 170% | 6.70% | 0.2393 | 0.2089 | 0.2049 | **0.2025** | 2.26 | 2.09 | **2.34** |
| 18 | 180% | 6.51% | 0.2484 | **0.1979** | 0.2061 | 0.1995 | **2.51** | 2.02 | 2.19 |
| 19 | 190% | 6.32% | 0.2572 | **0.1948** | 0.2010 | 0.1946 | **3.00** | 2.09 | 2.38 |
| 20 | 200% | 6.15% | 0.2658 | **0.1890** | 0.1997 | 0.1910 | **2.89** | 2.09 | 2.63 |
| 21 | 2000% | 1.02% | 0.7407 | **0.0949** | 0.1844 | 0.1384 | **4.68** | 2.11 | 2.91 |
| Panel C (1/$X$, $Y/X$, $Y$, and $X$ in final regression); $e$ = 10% of $X$ | | | | | | | | | |
| 22 | 20% | 6.02% | 0.0428 | 0.0498 | 0.0522 | 0.0509 | 0.86 | 0.79 | 0.76 |
| 23 | 30% | 5.71% | 0.0616 | 0.0562 | 0.0598 | 0.0579 | **1.06** | 0.88 | 0.94 |
| 24 | 40% | 5.44% | 0.0791 | 0.0647 | 0.0678 | 0.0657 | 1.22 | **1.25** | 1.21 |
| 25 | 50% | 5.19% | 0.0955 | 0.0748 | 0.0796 | 0.0770 | **1.38** | 1.25 | 1.22 |
| 26 | 60% | 4.96% | 0.1109 | 0.0835 | 0.0874 | 0.0846 | 1.46 | 1.41 | 1.44 |
| Panel D (1/$X$ and $Y/X$ in final regression); $e$ = 10% of $X$ | | | | | | | | | |
| 27 | 30% | 5.71% | 0.0616 | 0.1548 | 0.1541 | 0.1542 | 0.80 | 0.70 | 0.75 |
| 28 | 40% | 5.44% | 0.0791 | 0.1485 | 0.1478 | 0.1476 | 1.15 | 0.98 | 1.00 |
| 29 | 50% | 5.19% | 0.0955 | 0.1417 | 0.1441 | 0.1422 | 1.30 | 1.16 | 1.28 |
| 30 | 60% | 4.96% | 0.1109 | **0.1365** | 0.1417 | 0.1384 | **1.59** | 1.29 | 1.44 |
| 31 | 70% | 4.75% | 0.1254 | **0.1322** | 0.1409 | 0.1357 | **1.75** | 1.47 | 1.49 |
| 32 | 80% | 4.56% | 0.1392 | **0.1303** | 0.1439 | 0.1356 | **1.74** | 1.35 | 1.67 |
| 33 | 90% | 4.38% | 0.1523 | **0.1271** | 0.1438 | 0.1337 | **2.02** | 1.40 | 1.85 |
| 34 | 100% | 4.21% | 0.1648 | **0.1251** | 0.1491 | 0.1356 | **2.14** | 1.35 | 1.64 |

efficient observations. The opposite happens when $q$ makes a relatively small percent change in the true slope. In this case the dataset is flatter, causing the efficient observations to be more concentrated on the left and right and less concentrated on the top and bottom. When this happens (columns 1–3 of Tables 1 and 2), then LR-RTPLS is more accurate than its alternatives. In between the extremes of LR-RTPLS and UD-RTPLS is 4D-RTPLS which projects in all four directions and explains columns 4 and 5 of Tables 1 and 2. The presence of measurement and round-off error ($e$) makes it harder for [][]RTPLS to correctly capture the influence of the omitted variables. Error ($e$) also vertically shifts the frontier upwards. Thus, when $e$ gets larger, its influence is diminished by projecting left and right (LR-RTPLS). This explains line 3c of column 6 of Tables 1 and 2 as it compares to line 3d, columns 4 and 5.

Table 2 (comparing line 2 of Panels A and B) also shows that as the sample size increases from 100 observations to 500 observations, the standard deviation of [][]RTPLS error fell when $e$ is 0% (columns 1, 4, and 7) and when $q$ makes a 1000% difference and $e = 1\%$ (column 8). In all other cases, increasing the sample size caused the standard deviation of [][]RTPLS error to increase. In contrast, changing the sample size (or changing the amount of measurement and round off error) did not noticeably change the standard deviation of the error for OLS (Table 2, line 1). However, increasing the importance of $q$ does increase the standard deviation of the error for OLS. Furthermore OLS has a smaller standard deviation of the error than [][]RTPLS when $q = 10\%$ and $e = 1\%$ or 10% and when $q = 100\%$ and $e = 10\%$ for both sample sizes (Table 2, line 2, columns 2, 3, and 6). In all other cases, [][]RTPLS has a smaller standard deviation of the error than OLS. When the ratio between OLS and [][]RTPLS of the standard deviation of the error is found for each observation and then the mean is found (using the log procedure described above), OLS has a greater standard deviation of the error than [][]RTPLS for all cases; the OLS/[][]RTPLS ratio ranges from 1.07 to 1.55.

The patterns found in Tables 1 and 2 for the best projection direction are repeated in Panel B of Tables 3 and 4. Tables 3–5 use the same set of 100 values for $X$, $q$, and $\varepsilon$. Leightner and Inoue [5] generated the values for $X$, $q$, and $\varepsilon$ as random numbers between 0 and 10 and imposed no distributional assumptions (they also list the $X$ and $q$ data in their Table 1 and the $\varepsilon$ data in footnote 5 of Table 5). The dependent variable ($Y$) for Table 3 (both panels) was generated by plugging in the values for $X$, $q$, and $\varepsilon$ into $Y = 5 + X + \alpha Xq + 0.4\varepsilon$ where the numerical value for the $q$% given in Table 3, column 2, is 1000 times $\alpha$ and $0.4\varepsilon$ represents measurement and round-off error ($e$). Since both $X$ and $\varepsilon$ are series of numbers that range from 0 to 10, multiplying $\varepsilon$ by 0.4 makes $e$ equal to 40% of $X$.[12] The $e$% given in column 3 of Tables 3 and 4 is "$e$ as a percent of $Y$" and was calculated as the maximum value for $e$ divided by (the maximum value of $Y$ minus the maximum value for $e$). $Y$ for Table 4, Panels A and B, was calculated as $Y = 5 + X + \alpha Xq + 0.2\varepsilon$. Thus for these two panels, $e$ is 20% of $X$. Likewise the $Y$ of Table 4, Panels C and D, were calculated as $Y = 5 + X + \alpha Xq + 0.1\varepsilon$; thus $e = 10\%$ of $X$.

Each successive row of a given panel in Tables 3 and 4 represents an increase in the importance of $q$ as shown in column 2. The mean error and the OLS/[][]RTPLS ratios in Tables 3–5 were calculated in the same way as they were in Table 1, sans the taking of the mean value of 5000 simulations. Just as was done for Table 1, all the combinations of UD-RTPLS, LR-RTPLS, and 4D-RTPLS with three different sets of possible explanatory variables for the final regression were considered: $\{1/X, Y/X\}$, $\{1/X, Y/X, Y\}$, and $\{1/X, Y/X, Y, X\}$. For Table 3, Panel A, and for Table 4, Panels A and C, the best set of explanatory variables for the final regression was always $1/X$, $Y/X$, $Y$, and $X$ (and only those results are presented). Likewise, for Table 3, Panel B, and for Table 4, Panels B and D, the best set of explanatory variables for the final regression was always $1/X$ and $Y/X$ (and only those results are presented). These patterns mirror the patterns found in Table 1 where $1/X$, $Y/X$, $Y$, and $X$

**Table 5:** Other specifications.

| Estimated | True equation | \|Mean error\| | | | OLS/RTPLS | OLS/[][]RTPLS | Projection[1] |
|---|---|---|---|---|---|---|---|
| | | OLS | RTPLS | [][]RTPLS | | | |
| (1) $Y = \alpha_0 - \alpha_1 X$ | (a) $Y = 101 - X - Xq$ | 0.594 | 0.533 (0.115) | 0.085 | 1.260 (5.866) | 3.893 | UD+ (1D+) |
| | (b) $Y = 101 - X - 0.1Xq$ | 0.164 | 0.158 (0.020) | 0.022 | 1.251 (8.920) | 5.701 | UD (1D) |
| | (c) $Y = 101 - X - 0.01Xq$ | 0.022 | 0.023 (0.003) | 0.007 | 1.200 (6.989) | 4.063 | UD+ (1D+) |
| | (d) $Y = 101 - X + Xq$ | 2.294 | 0.475 (0.206) | 0.159 | 5.655 (4.985) | 5.741 | UD (1D) |
| | (e) $Y = 101 - X + 0.1Xq$ | 1.338 | 0.884 (0.091) | 0.013 | 1.024 (6.573) | 44.188 | LR (1D++) |
| | (f) $Y = 101 - X + 0.01Xq$ | 0.024 | 0.024 (0.013) | 0.009 | 1.166 (2.293) | 2.119 | LR+ (1D) |
| | (g) $Y = 101 - X + 0.2Xq$ | 1.239 | 1.562 (0.452) | 0.646 | 1.681 (2.330) | 1.507 | UD (1D) |
| | (h) $Y = 101 - X + 0.3Xq$ | 4.490 | 4.607 (0.119) | 0.117 | 1.725 (40.254) | 26.941 | UD (1D) |
| | (i) $Y = 101 - X + 0.4Xq$ | 4.739 | 2.860 (0.246) | 0.215 | 2.210 (9.029) | 10.469 | UD (1D) |
| | (j) $Y = 101 - X + 0.5Xq$ | 4.868 | 2.626 (1.107) | 0.972 | 2.226 (6.605) | 9.150 | UD (1D) |
| (2) $Y = \alpha_0 + \alpha_1 X$ | (a) $Y = 5 + X^2 + Xq$ | 0.368 | 0.340 (0.350) | 0.243 | 1.015 (2.294) | 3.018 | LR (1D+) |
| $Y = \alpha_0 + \alpha_1 X^2$ | (b) $Y = 5 + X^2 + X^2 q$ | 0.566 | 0.225 (0.109) | 0.018 | 1.619 (3.641) | 47.400 | 4D (1D+) |
| (3) $^2\mathrm{Ln}(Y) = \alpha_0 + \alpha_1 \mathrm{Ln}(X)$ | (a) $Y = e^5 X^{1+q}$ | 0.636 | 0.652 (0.302) | 0.242 | 0.903 (1.026) | 1.223 | UD (1D) |
| | (b) $Y = e^5 X^{1+0.1q}$ | 0.172 | 0.043 (0.166) | 0.134 | 2.684 (1.244) | 1.500 | UD++ (1D++) |
| | (c) $Y = e^5 X^{1+0.01q}$ | 0.023 | 0.009 (0.007) | 0.011 | 2.333 (2.483) | 1.500 | UD (1D+) |
| (4) $Y = \alpha_0 + \alpha_1 X$ | (a) $Y = 5 + X + Xq$ | 0.594 | 0.152 (0.103) | 0.106 | 4.424 (8.473) | 4.660 | BD+ (1D+) |
| | (b) $Y = 5 + X + X^2 q$ | 3.410 | 1.432 (1.583) | 1.176 | 1.457 (3.020) | 2.926 | LR (1D++) |
| | (c) $Y = 5 + X + X^3 q$ | 27.38 | 9.478 (14.19) | 9.349 | 1.675 (3.380) | 4.240 | 4D+ (1D+) |
| (5) $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2$ | (a) $Y = 5 + X_1 + X_2 + X_1 q$ | 0.593 | 0.081 (0.092) | 0.090 | 5.65 (16.29) | 9.686 | UD+ (1D+) |
| | (b) $Y = 5 + X_1 + X_2 + 0.1X_1 q$ | 0.164 | 0.086 (0.040) | 0.020 | 1.425 (3.936) | 15.126 | UD (1D+) |
| | (c) $Y = 5 + X_1 + X_2 + 0.01X_1 q$ | 0.022 | 0.011 (0.005) | 0.004 | 2.130 (5.985) | 5.661 | UD (1D) |

[1]1D = RTPLS; 4D = 4D-RTPLS; LR = LR-RTPLS; UD = UD-RTPLS. No +, +, ++ = $\{1/X, Y/X\}$, $\{1/X, Y/X, Y\}$, and $\{1/X, Y/X, Y, X\}$ used as the explanatory variables in the final regression, respectively.

were the best explanatory variables in column 3 and $1/X$ and $Y/X$ were the best explanatory variables in all other columns. Notice that Panels B and D are extensions of Panels A and C, respectively, with several rows of overlap presented (see the $q$% given in column 2).

In Table 3 (where $e$ = 40% of $X$), LR-RTPLS, 4D-RTPLS, and BD-RTPLS produced the largest OLS/[][]RTPLS ratio when $q$ affected the true slope by 300% to 380%, 390% to 440%, and more than 440%, respectively. This progression from LR-RTPLS to 4D-RTPLS to BD-RTPLS as $q$ increases in importance reflects the progression shown in Table 1. Furthermore, it is reflected in Table 4, Panel B. In Table 4 (where $e$ = 20% of $X$), LR-RTPLS, 4D-RTPLS, and BD-RTPLS produced the largest OLS/[][]RTPLS ratio when $q$ affected the true slope by 120% to 140%, 150% to 170%, and more than 170%, respectively. Thus a smaller amount of $e$ (Table 4, Panel B) leads to narrower ranges for LR-RTPLS and 4D-RTPLS at much smaller values for the importance of $q$ than did the case with a larger amount of $e$ in Table 3, Panel B. In Table 4, Panels C and D, $e$ as a percent of $X$ falls even more (to 10%) and the results show no region (given our increasing the importance of $q$ by 10% for each row), where LR-RTPLS and 4D-RTPLS are best.

Finally, notice that the mean of the absolute value of OLS's error always increases as the importance of $q$ **increases** (column 4 of Tables 3 and 4); in contrast the mean of the absolute value of BP-RTPLS's error always **falls** (columns 5–7) when estimates using just $1/X$ and $Y/X$ are optimal (Panels B and D). In all the cases shown in Tables 3 and 4, if $e$ is less than 5% of $Y$, then UD-RTPLS using $1/X$ and $Y/X$ in the final regression is the BP-RTPLS method.

Table 5 replicates the results of Table 5 of Leightner and Inoue [5] for applying the first generation of this technique (RTPLS) to different types of equations and compares those results to BP-RTPLS. Column 1 gives the equation estimated. Column 2 gives the true equation into which the data from Tables 1 and 5 of Leightner and Inoue [5] was inserted. Table 5, column 3, presents the mean of the absolute value of the error for OLS (calculated using (4.6), sans the taking of the mean of 5000 simulations). Column 5 gives the mean of the absolute value of the error for BP-RTPLS, column 7 gives the OLS/BP-RTPLS ratios, and column 8 tells what specific form BP-RTPLS took—UD, LR, 4D correspond to UD-RTPLS, LR-RTPLS, and 4D-RTPLS, respectively; no + signs, one + sign, and two + signs after UD, LR, and 4D indicate $\{1/X, Y/X\}$, $\{1/X, Y/X, Y\}$, and $\{1/X, Y/X, Y, X\}$ as the explanatory variables in the final regression, respectively. "1D" in column 8 denotes RTPLS.

The number not in parenthesis in columns 4 and 6 duplicates the numbers given in Table 5 of Leightner and Inoue [5] for the first generation of this technique (RTPLS) for the mean of the absolute value of the error for RTPLS and for the OLS/RTPLS ratio. The numbers in parenthesis in columns 4 and 6 show how RTPLS would have performed if a constant had not been included in the final regression.[13] A comparison of the numbers not in parenthesis to those in parenthesis dramatically illustrates how important it is to not include a constant in the final regression—not including a constant increased the OLS/RTPLS ratio for all but two of the cases (lines 1d and 3b) and the average OLS/RTPLS ratio increased 3.82-fold.[14]

If $\partial Y/\partial X$ might be negative (Line 1, Table 5), then a preliminary OLS regression should be run between $X$ and $Y$. If this preliminary regression generates a positive $dY/dX$ (as it did for lines 1d, 1g, 1h, 1i, and 1j), then normal BP-RTPLS can be used (note: true $\partial Y/\partial X$ was negative for 4, 43, 26, 20, and 16 percent of the observations in lines 1(d), 1(g), 1(h), 1(i), and 1(j), resp.). However, the preliminary regression found a negative $dY/dX$ for the cases given in lines 1(a), 1(b), 1(c), 1(e), and 1(f). In these cases, all $Y$s were multiplied by negative one and then a constant (equal to 101, which was sufficiently big to make all $Y$s positive) was added to all $Y$s. The normal BP-RTPLS process was then conducted using the adjusted

$Ys$, but the resulting $\partial Y/\partial Xs$ were remultiplied by minus one. Multiplying either $Y$ or $X$ by negative one and then adding a constant to make them all positive is necessary because (2.7) and (2.8) only work for positive relationships.

This entire paper deals with misspecification error in that the influence of omitted variables is ignored when using OLS for all of this paper's cases. However, Table 5, line 2(a) takes misspecification error to even the relationship between $Y$ and $X$: $X$ should be squared (column 2), but it is not (column 1). In this case BP-RTPLS produced 24 percent mean error (column 5) and a third of the error of OLS (column 7). Line 3 shows the results of using RTPLS when omitted variables affect an exponent. Line 4 of Table 5 demonstrates that the relationship between the omitted variable and the known independent variable does not have to be modeled for BP-RTPLS to work well; BP-RTPLS noticeably out performs OLS when the interaction term is $X_1q$ (Line 4(a)), $X_1^2q$ (Line 4(b)), and $X_1^3q$ (Line 4(c)).

Line 5 of Table 5 shows how BP-RTPLS can be used when there is more than one known independent variable, where only one of them interacts with omitted variables. Leightner and Inoue [5] argue that OLS produces consistent estimates for the known independent variables that do not interact with omitted variables. Therefore to apply BP-RTPLS to the equation in Line 5 of Table 5, an OLS estimate can be made of $Y = \alpha_0 + \alpha_1X_1 + \alpha_2X_2$. $Y_1$ can then be calculated as $Y - \alpha_2\hat{}X_2$. Finally RTPLS can be used normally to find the relationship between $Y_1$ and $X_1$ (note: in Table 5, Line 5 the error from OLS is from estimating $Y = \alpha_0 + \alpha_1X_1 + \alpha_2X_2$). In all the cases shown in Table 5, BP-RTPLS noticeably out performs OLS. Comparing column 7 to column 6 of Table 5 and line 3(a) to 3(b) of Table 1 clearly shows that BP-RTPLS produces a major improvement over the first two generations of this technique.

## 5. Example

When the government buys goods and services (G), it causes gross domestic product (GDP) to increase by a multiple of the spending. The pathways linking G and GDP are numerous, interacting, and complex. For example, the increased government spending will cause producer and consumer incomes to rise, interest rates to rise, and put upward or downward pressure on the exchange rate, affecting exports and imports which in turn affect GDP. Many economists have spent their careers trying to model all the important interconnections in order to better advise the government. To complement the efforts of these economists, BP-RTPLS can be used to produce reduced form estimates of ∂GDP/∂G without having to model all the "omitted variables."

Annual data for the USA between 1929 and 2010 were downloaded from the Bureau of Economic Analysis Website (http://www.bea.gov/). The data were in billions of 2005 dollars and corrected for inflation using a chain-linked index method. The top line of Figure 3 shows the results of using LR-RTPLS and the bottom line of using UD-RTPLS to estimate ∂GDP/∂G. If 4D-RTPLS had been depicted, it would lie between the top and bottom lines. Although LR-RTPLS and UD-RTPLS produced different estimates, the two lines are close to each other and they are approximately parallel.

The UD-RTPLS (LR-RTPLS) ∂GDP/∂G estimate for 2010 of 6.01 (6.26) implies that a one dollar increase in real government spending would cause real GDP to increase by 6.01 (6.26) dollars. The big dip down in ∂GDP/∂G coincides with WWII—the UD-RTPLS (LR-RTPLS) estimate of ∂GDP/∂G in 1940 was 5.44 (5.67) and it fell to 1.65 (1.73) in 1945. It makes sense that the government purchasing bullets, tanks, and submarines (many of which were destroyed in WWII) would have a smaller multiplier effect than the government building

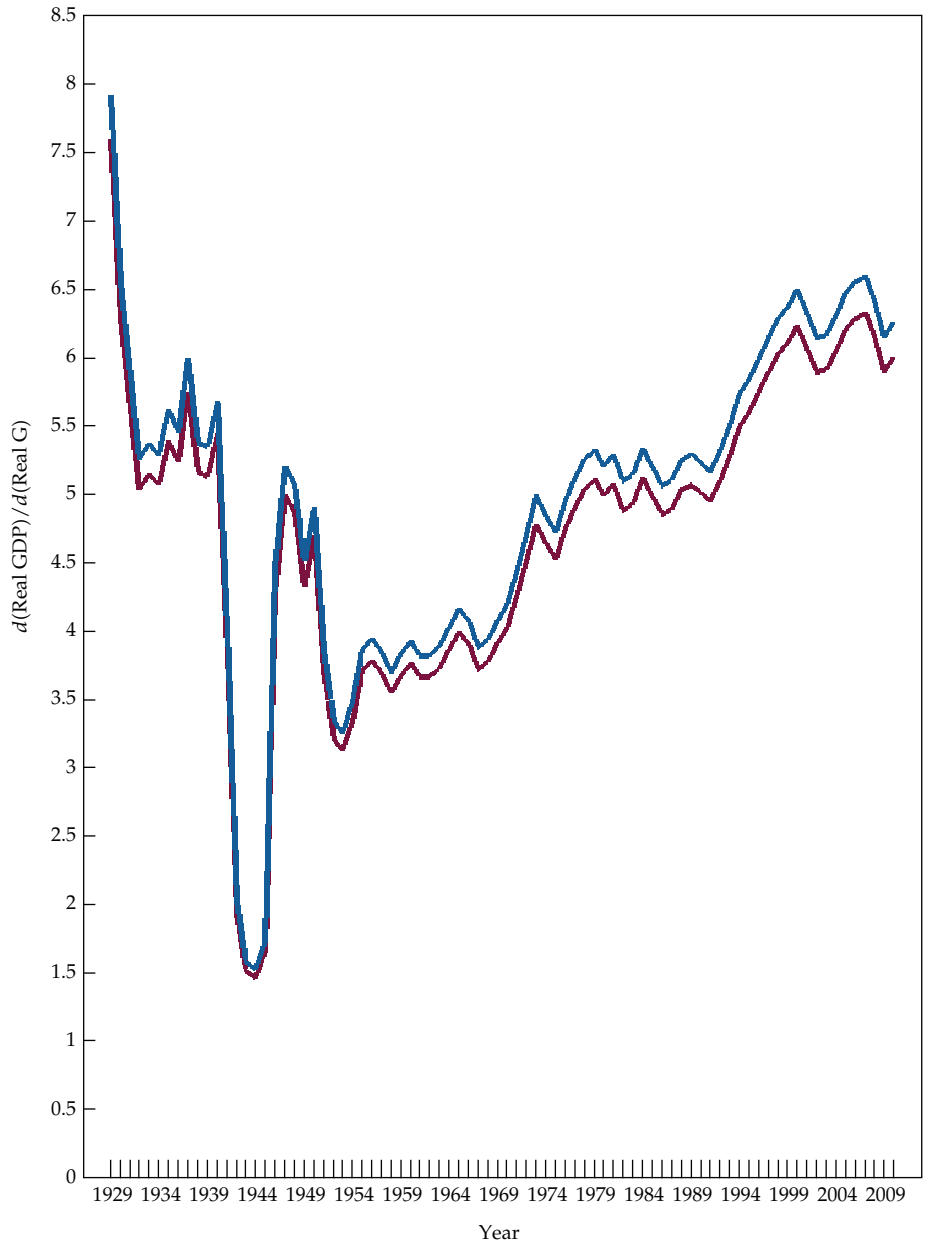



**Figure 3:** $d$(Real GDP)/$d$(Real G) for the USA (Top line = LR-RTPLS; Bottom Line = UD-RTPLS).

roads and schools during nonwar times. The UD-RTPLS (LR-RTPLS) estimates climbed from 3.12 (3.26) in 1953 to 6.33 (6.60) in 2007. The crisis that started in the USA in 2008 caused the government multiplier to fall by five percent. An OLS estimate of $\partial \text{GDP}/\partial \text{G}$ is 5.22 for all years.

## 6. Conclusion

This paper has developed and extensively tested a third generation of a technique that uses the relative vertical position of observations to account for the influence of omitted variables that interact with the included variables without having to make the strong assumptions of proxies or instruments. The contributions of this paper include the following.

First, Leightner and Inoue [5] showed that RTPLS has less bias than OLS when there are omitted variables that interact with the included variables. However, this paper shows that both RTPLS and BD-RTPLS (the first two generations of this technique) still contained some bias (see footnote 9) because it included a constant in the final regression. Section 3 of this paper shows that the third generation of this technique (BP-RTPLS) is not biased. Second, this paper shows that when [][]RTPLS does not include a constant, it produced OLS/[][]RTPLS ratios that were 586 percent higher on average than [][]RTPLS when it does include a constant in Table 1 (ignoring column 3) and 382 percent higher in Table 5. Deleting this constant constitutes a major improvement.

Second, this is the first paper to test how the direction of data projection and the variables included in the final regression affect the results. Very strong and persistent patterns were found that include (1) that $1/X$, $Y/X$, $Y$, and $X$ should be used as the explanatory variables in the final regression when $q$ has an extremely small effect on the true slope and that only $Y/X$ and $1/X$ should be used when $q$ has a normal or relatively larger effect on the true slope[15], (2) as the importance of the omitted variable increases, and as the size of measurement and round off error decreases, there is usually a range where LR-RTPLS produces the best estimates followed by a range where 4D-RTPLS is best, followed by UD-RTPLS being best. However, UD-RTPLS using just $1/X$ and $Y/X$ in the final regression will be (by far) the best procedure for the widest range of possible values for the importance of $q$, for the size of $e$, and for the type of specification. We recommend that researchers wanting to use BP-RTPLS use UD-RTPLS but test the robustness of their results by comparing them to (at the very least) LR-RTPLS estimates and then focus their analysis on conclusions that can be drawn from both the UD-RTPLS and LR-RTPLS estimates.

## Acknowledgments

The authors appreciate the comments and suggestions made on earlier generations of BP-RTPLS by Knox Lovell, Ron Smith, Lawrence Marsh, and Charles Horioka.

## Endnotes

1. If the true relationship is $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2$ and if $X_2$ is omitted from the analysis, but $X_2$ has no relationship with $X_1$ and thus does not affect the true slope—$dY/dX_1$—then $X_2$ acts as a source of additional random variation which (in large samples) does not change the numerical value of the estimated slope ($\alpha_1$); however, it will affect the estimated level of statistical significance. One indicator of the importance of "omitted variable bias" is that a Google Scholar search conducted in September 2011 generated 276,000 hits for that phrase. Those hits included [16–30]. These papers include applications ranging from criminal arrest rates, school achievement, hospital costs, psychological distress, housing values, employment, health care expenditures, the cost of equity capital, effects of unemployment insurance, productivity, and financial aid for higher education.

2. The estimate for $\partial Y/\partial X$ will be approximately $\alpha_1 + \alpha_2 E(q^m)$, where $E(q^m)$ is the expected, or mean, value for $q^m$.

3. Instrumental variables must also be ignorable, or not add any explanatory value independent of their correlation with the omitted variable. Furthermore, they must be so highly correlated with the omitted variable that they capture the entire effect of the omitted variable on the dependent variable [1]. Other methods for addressing omitted variable bias (e.g., see [20, 22, 28, 30]) also require questionable assumptions that are not made by BP-RTPLS.

4. If, instead of adding $0.4qX$ in (1.1), we had subtracted $0.4qX$, then the smallest $qs$ would be on the top and the largest $qs$ on the bottom of Figure 1. Either way, the vertical position of observations captures the influence of the omitted variable $q$.

5. We say "approximately" because how the data are projected to the frontier will affect the resulting TPLS estimate. If the data are projected upwards, then the top of the frontier is weighted heavier. If the data are projected to the left, then the bottom of the frontier is weighted heavier. Notice that projecting to the upper frontier in Figure 1 eliminated approximately 92 percent of the variation due to omitted variables (the $qs$ was changed from a range from 1 to 98 to a range from 92 to 98). The final regression eliminates any remaining variation due to omitted variables.

6. $Y$ is more likely than $X$ to be correlated to $n$; thus we consider adding either $Y$ or $Y$ and $X$, but not just $X$. Notice that (2.6) should be estimated without using a constant.

7. All of the existing RTPLS and BD-RTPLS literature truncated off any horizontal and vertical regions of the frontier, truncated off 3% of the other side of the frontier, and used a constant in the final regression. BP-RTPLS does not truncate off 3% of the other side of the frontier nor does it add a constant to the final regression.

8. Think of a straight regression line that would pass through the observations on the frontier. The slope of that regression line would be flatter than the slope through just the 90 s and steeper than the slope going through all the 80 s in Figure 2. Also notice that in this case the second iteration will return to producing a perfect slope estimate for the remaining observations associated with a $q$ of 80, after truncating off a small horizontal region of the frontier.

9. If we plotted the true value of the slope versus the BP-RTPLS estimate of the slope, then BP-RTPLS works perfectly if its estimates lie on the 45 degree line. The effect discussed in this paragraph implies that if a constant was added to the final BP-RTPLS estimate (which would be incorrect), then the BP-RTPLS line would cross the 45 degree line in the middle of the data and the triangle formed by the 45 degree line and the BP-RTPLS line below this crossing would be identical to the triangle formed above this crossing. This is exactly what we find if we add a constant to the final BP-RTPLS estimate. However, when a constant is not included, the two triangles being of equal size off set each other and the BP-RTPLS estimates lie along the 45 degree line indicating the absence of bias. This implies that adding a constant to the final regression, as was done in the first two generations of this technique, resulted in biased estimates; however, this is not a problem in the third generation.

10. This BD-RTPLS is not exactly the same as the second generation of this technique. It is like the second generation in that it peels the data both down and up and that it used a constant, $1/X$ and $Y/X$ in the final regression. It is unlike the second generation because

it did not truncate off the smallest 3% of the $Xs$ in each iteration when peeling down and the largest 3% of the $Xs$ when peeling up. However, by making the difference between BD-RTPLS in line 2a and UD-RTPLS in line 2b solely the presence of a constant in the final regression for BD-RTPLS, we dramatically illustrate why a constant should not be included in the final regression. To see this, compare line 3(a) and 3(b).

11. Leightner and Inoue [5] mistakenly substituted "Ln($|E_i^{\text{ols}}|/|E_i^{\text{RTPLS}}|$)" for $E_i^{\text{ols}}$ in their counterpart for (4.6). This resulted in the absolute value being taken twice, which should not have been done. This affected their results the most when the size of measurement error was 10%.

12. Thus this $e$ is always positive. This always positive $e$ can be thought of as the combined effects of an omitted variable that shifts the relationship between $Y$ and $X$ upwards (without changing its slope) with measurement and round-off error that would sometimes increase and sometimes decrease $Y$. This was the easiest way to construct error that can be calibrated to $X$.

13. The old RTPLS not only used a constant in the final regression, it also truncated off the first 3% of the frontier which occurred on the side of the frontier opposite any potentially horizontal or vertical region and it did not make estimates for the observations that corresponded to the 3% of the observations with the smallest values for $X$. The numbers given in parentheses in columns 4 and 6 of Table 5 do none of these things. However, the numbers in parenthesis do use the best set of explanatory variables for the final regression: $\{1/X, Y/X\}$, $\{1/X, Y/X, Y\}$, or $\{1/X, Y/X, Y, X\}$ as indicated in column 8.

14. For approximately half the cases, BP-RTPLS estimates (column 7) were less than the RTPLS estimates when a constant is not used (numbers in parentheses in column 6). This implies that the TPLS estimates from peeling the data downward were more accurate than the TPLS estimates from peeling the data upwards for this dataset.

15. Table 5 shows that this rule may not hold for other specifications. Much more work needs to be done to determine the optimal set of explanatory variables for the final regression under different specifications.

## References

[1] K. A. Clarke, "The phantom menace: omitted variable bias in econometric research," *Conflict Management and Peace Science*, vol. 22, no. 4, pp. 341–352, 2005.

[2] K. A. Clarke, "Return of the phantom menace: omitted variable bias in political research," *Conflict Management and Peace Science*, vol. 26, no. 1, pp. 46–66, 2009.

[3] J. Branson and C. A. K. Lovell, "Taxation and economic growth in New Zealand," in *Taxation and the Limits of Government*, G. W. Scully and P. J. Caragata, Eds., pp. 37–88, Kluwer Academic, Boston, Mass, USA, 2000.

[4] J. E. Leightner, *The Changing Effectiveness of Key Policy Tools in Thailand*, Institute of Southeast Asian Studies for East Asian Development Network, 2002, EADN Working Paper 19(2002) x0219-6417.

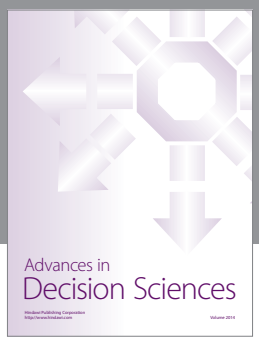
[5] J. E. Leightner and T. Inoue, "Tackling the omitted variables problem without the strong assumptions of proxies," *European Journal of Operational Research*, vol. 178, no. 3, pp. 819–840, 2007.

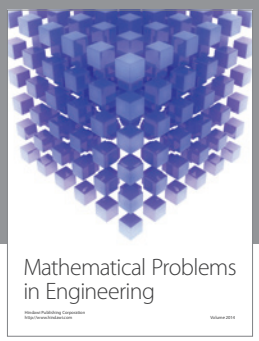
[6] J. E. Leightner, "Omitted variables and how the Chinese yuan affects other Asian currencies," *International Journal of Contemporary Mathematical Sciences*, vol. 3, no. 13-16, pp. 645–666, 2008.

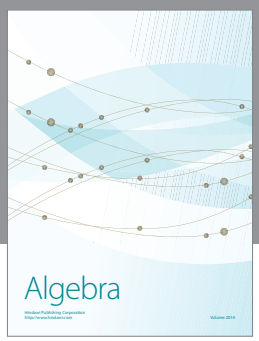
[7] J. E. Leightner, "China's fiscal stimulus package for the current international crisis: what does 1996–2006 tell us?" *Frontiers of Economics in China*, vol. 5, no. 1, pp. 1–24, 2010.

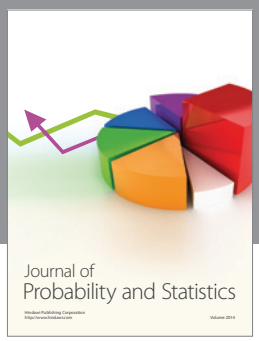
[8] J. E. Leightner, "Fiscal stimulus for the USA in the current financial crisis: what does 1930–2008 tell us?" *Applied Economics Letters*, vol. 18, no. 6, pp. 539–549, 2011.

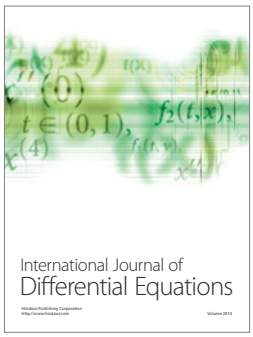
[9] J. E. Leightner, "How China's holdings of foreign reserves affect the value of the US dollar in Europe and Asia," *China & World Economy*, vol. 18, no. 3, pp. 24–39, 2010.
[10] J. E. Leightner, "Omitted variables, confidence intervals, and the productivity of exchange rates," *Pacific Economic Review*, vol. 12, no. 1, pp. 15–45, 2007.
[11] J. E. Leightner, "Fight deflation with deflation, not with Monetary policy," *The Japanese Economy*, vol. 33, no. 2, pp. 67–93, 2005.
[12] J. E. Leightner, "The productivity of government spending in Asia: 1983–2000," *Journal of Productivity Analysis*, vol. 23, no. 1, pp. 33–46, 2005.
[13] J. E. Leightner and T. Inoue, "Negative fiscal multipliers exceed positive multipliers during Japanese deflation," *Applied Economics Letters*, vol. 16, no. 15, pp. 1523–1527, 2009.
[14] J. E. Leightner and T. Inoue, "Capturing climate's effect on pollution abatement with an improved solution to the omitted variables problem," *European Journal of Operational Research*, vol. 191, no. 2, pp. 539–556, 2008.
[15] J. E. Leightner and T. Inoue, "The effect of the Chinese Yuan on other Asian Currencies during the 1997-1998 Asian Crisis," *International Journal of Economic Issues*, vol. 1, no. 1, pp. 11–24, 2008.
[16] J. K. Abbott and H. A. Klaiber, "An embarrassment of riches: confronting omitted variable bias and multiscale capitalization in hedonic price models," *The Review of Economics and Statistics*, vol. 93, no. 4, pp. 1331–1342, 2011.
[17] J. D. Angrist and A. B. Krueger, "Instrumental variables and the search for identification: from supply and demand to natural experiments," *Journal of Economic Perspectives*, vol. 15, no. 4, pp. 69–85, 2001.
[18] S. E. Black and L. M. Lynch, "How to compete: the impact of workplace practices and information technology on productivity," *The Review of Economics and Statistics*, vol. 83, no. 3, pp. 434–445, 2001.
[19] C. A. Botosan and M. A. Plumlee, "A re-examination of disclosure level and the expected cost of equity capital," *Journal of Accounting Research*, vol. 40, no. 1, pp. 21–40, 2002.
[20] S. R. Cellini, "Causal inference and omitted variable bias in financial aid research: assessing solutions," *Review of Higher Education*, vol. 31, no. 3, pp. 329–354, 2008.
[21] P. Y. Crémieux and P. Ouellette, "Omitted variable bias and hospital costs," *Journal of Health Economics*, vol. 20, no. 2, pp. 271–282, 2001.
[22] T. A. DiPrete and M. Gangl, *Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Estimation With Imperfect Instruments*, Wissenschaftszentrum Berlin fur Sozialforschung, Berlin, Germany, 2004.
[23] A. L. Harris and K. Robinson, "Schooling behaviors or prior skills? A cautionary tale of omitted variable bias within oppositional culture theory," *Sociology of Education*, vol. 80, no. 2, pp. 139–157, 2007.
[24] D. B. Mustard, "Reexamining criminal behavior: the importance of omitted variable bias," *The Review of Economics and Statistics*, vol. 85, no. 1, pp. 205–211, 2003.
[25] R. K. Pace and J. P. LeSage, "Omitted variable biases of OLS and spatial lag models," in *Progress in Spatial Analysis*, Advances in Spatial Science, part 1, pp. 17–28, 2010.
[26] R. W. Paterson and K. J. Boyle, "Out of sight, out of mind? Using GIS to incorporate visibility in hedonic property value models," *Land Economics*, vol. 78, no. 3, pp. 417–425, 2002.
[27] R. M. Scheffler, T. T. Brown, and J. K. Rice, "The role of social capital in reducing non-specific psychological distress: the importance of controlling for omitted variable bias," *Social Science and Medicine*, vol. 65, no. 4, pp. 842–854, 2007.
[28] D. N. Sessions and L. K. Stevans, "Investigating omitted variable bias in regression parameter estimation: a genetic algorithm approach," *Computational Statistics & Data Analysis*, vol. 50, no. 10, pp. 2835–2854, 2006.
[29] S. C. Stearns and E. C. Norton, "Time to include time to death? The future of health care expenditure predictions," *Health Economics*, vol. 13, no. 4, pp. 315–327, 2004.
[30] P. A. V. B. Swamy, I. L. Chang, J. S. Mehta, and G. S. Tavlas, "Correcting for omitted-variable and measurement-error bias in autoregressive model estimation with panel data," *Computational Economics*, vol. 22, no. 2-3, pp. 225–253, 2003.

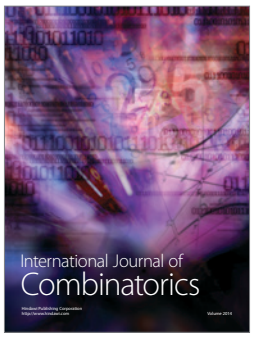


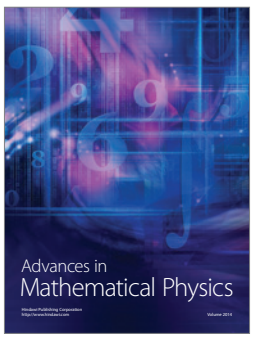


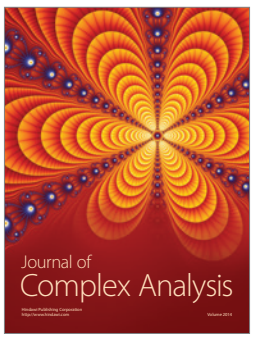


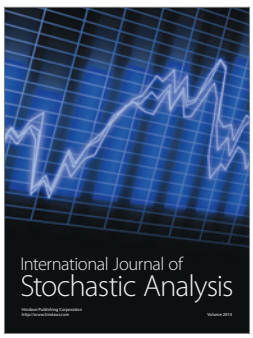


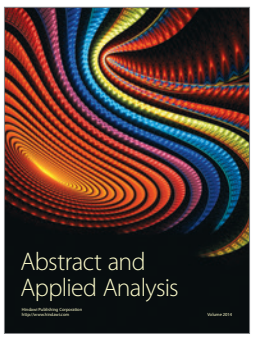


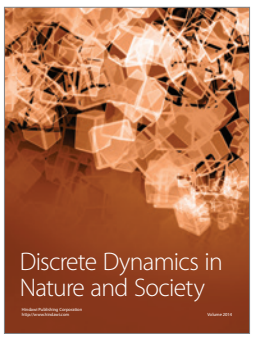


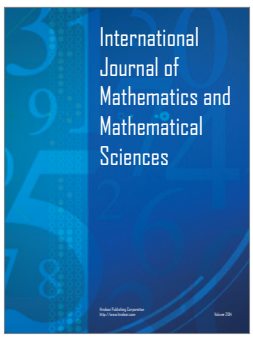


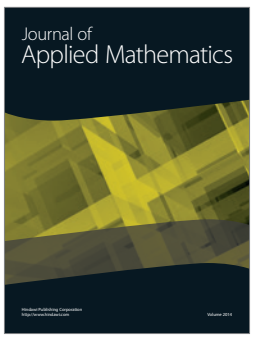












Hindawi
Submit your manuscripts at
http://www.hindawi.com