International Review of Financial Analysis 97 (2025) 103864

Contents lists available at ScienceDirect

International Review of Financial Analysis

journal homepage: www.elsevier.com/locate/irfa

A gentle reminder: Should returns be interpreted as log differences?

David Iheke Okorie [a,b,c,d,*]

[a] Hangzhou City University (HZCU), Hangzhou, China
[b] University of Waikato, Management School, School of Accounting, Fince and Economics, Hamilton, New Zealand
[c] Centre for the Study of the Economies of Africa (CSEA), Nigeria
[d] SD Consulting Agency (SCA), Nigeria

ARTICLE INFO

Keywords:
Return
Log difference
Approximation, Taylor expansion
Log(X + 1)
Simulation

ABSTRACT

It is rather a norm for researchers to directly use the log difference of an asset price to compute returns. Just like using $ln(X+1)$ to avoid taking the natural logarithm of zero(s). However, this log returns is but a conditional approximation of the actual returns. Nonetheless, can log difference approximations and the $ln(X+1)$ common practices produce BLUE estimates? Using the log return as an example, this study discusses the approximation nature and conditions for using the log difference approximation both for the interest regressor and control variables. These conditions are; that both the sample average and variance of the original series tend to zero. When these conditions are not met, the log difference approximation is, in fact, not a good approximation and biases OLS causal estimators. When the conditions are met, it produces unbiased, consistent but less efficient estimators. Thereby making the estimates less precise and less accurate. Nonetheless, this is true for a log differenced interest regressor(s) and control variables, when it correlates with the interest variable(s) and explains, in part, the dependent variable, even in large samples. Similarly, the common use of $ln(X+1)$ biases the estimation of the true causal effect, even the intercept term, except when X tends to infinity. A robust solution of using non-zero subsamples, against $ln(X+1)$, produces unbiased and consistent estimators for the true causal effects under the causal assumptions. These biasedness, inconsistencies, and inefficiencies do not disappear in large samples. Finally, both ex-ante and ex-post test statistics are discussed, however, the ex-post estimation test statistic is recommended to confirm both the choice of using log difference approximation and that of using $ln(X+1)$, in an empirical data causal regression analysis. Ideally, researchers should ensure the conditions for using the log difference approximation are met. Otherwise, these approximations and practices produce biased, inconsistent, and inefficient results, even in large samples, leading to misinformed policy implications.

1. Introduction

If the return of an asset is 5 %, the logarithmic approximation of return (log return) will have an error of 0.121 percentage points. If the return of an asset is 10 %, the log return (log difference) approximation will have an error of 0.469 percentage points. On the other hand, if the return of an asset is −5 % (−10 %), the log return will have an error of 0.129 (0.536). This shows the asymmetric nature of the logarithmic approximation errors. That is, the log return errors for negative returns outweigh those for positive returns. These errors increase in the asset's return, that is, the errors increase as the asset's returns increase. For asset returns, $r_t \in [0, 0.5]$, Fig. 1 shows the error of log return, x_t. Nonetheless, these errors are not the only problems applied econometricians and/or empirical researchers face. A more serious problem is that log returns, instead of using the actual returns (Okorie, 2023; Okorie, Bouri, & Mazur, 2024; Okorie & Lin, 2023), can render the OLS (Ordinary Least Squares) estimators biased, inconsistent, and inefficient. This problem intensifies when the sample average and variance of the asset's returns do not tend to zero. Beyond this, the bias, inconsistency, and inefficiency of the log return OLS estimator get worse in high-variance asset return samples and the presence of heteroscedasticity. Besides, even when the asset returns' sample mean and variance tend to zero, the OLS estimators of log returns are less efficient, that is the OLS estimators might not be BLUE (Best Linear Unbiased Estimator). These

Abbreviations: BLUE, Best Linear Unbiased Estimator; OLS, Ordinary Least Squares; OVB, Omitted Variable Bias; CLT, Central Limits Theory; LLN, Law of Large Number; DGF, Data Generating Function; OV, Omitted Variable; CV, Control Variable.

* Corresponding author at: Hangzhou City University (HZCU), Hangzhou, China.
E-mail address: okorie.davidiheke@gmail.com.

https://doi.org/10.1016/j.irfa.2024.103864
Received 30 August 2024; Received in revised form 1 December 2024; Accepted 4 December 2024
Available online 9 December 2024

issues of log approximation are illustrated in this study using log return as an example.

Log return or log differencing of prices, an example of log approximation, is one of the common practices in empirical studies. Reexamining common practices or norms is always a vital advancement strategy. In fact, without these reevaluations, Halvorsen and Palmquist (1980) would not have discovered the long-standing misinterpretation of semilogarithmic log-linear models with dummy regressors. After their paper, researchers began to interpret their semilogarithmic models with caution. Efforts are still being made to address other common practices (Bellemare & Wichman, 2019; Mullahy & Norton, 2023; Nick, 2023). In the same light, this paper seeks to call the attention of researchers to re-examine and re-evaluate the use of log-return or log-price-difference approximation for asset returns. This is suicidal and distorts valid inference when the conditions for using log-difference approximation are not met. Thus, log-differencing an asset's price can produce misleading results and lead to invalid inferences, even in large samples. In the past, most studies have ignored the conditions for log approximations. Examples may include, but are not limited to, Duarte-Silva and Kimel (2024), Panagiotidis, Papapanagiotou, and Stengos (2024), Long, Chiah, Zaremba, and Umar (2024), Liu and Kang (2024), Wong (2023), Domenico, Livan, Montagna, and Nicrosini (2023), Blau, Griffith, and Whitby (2023), Chiang and Chen (2023), Simonato and Denault (2023), Ni and Wang (2023), Ausloos, Ficcadenti, Dhesi, and Shakeel (2021), etc. On the other hand, some studies like Tomlinson, Greenwood, and Mucha-Kruczyński (2024), and Herley, Orlowski, and Ritter (2023), used log return and showed that the mean and variance of the log returns are sufficiently close to zero.

However, there are still a few concerns. Firstly, the mean and variance of the log return being sufficiently close to zero are not the conditions for using the log return approximation. Secondly, log returns always under-predict or under-approximate the actual returns. As such, having the mean and variance of the log return close to zero does not necessarily imply that the mean and variance of the actual return (not the log return) are sufficiently close to zero (Okorie, Gnatchiglo, & Wesseh, 2024); which are the conditions to approximate returns with log returns. Nonetheless, the OLS estimates of the log return approximations are relatively inefficient (less precise and less accurate) compared to that of the actual return even when these conditions are met. Another common research practice, which is not addressed in this current study, is to add one to a variable and take the natural log of the sum (Dong & Yu, 2023; Fang, Tian, & Tice, 2014; Liu & Kang, 2024; Pungaliya & Wang, 2023; Zhang, 2022). Some researchers claim to follow published papers from top journals, such as Fang et al. (2014), in this practice while others claim that this is done to be able to take the natural logarithmic value of zero. The challenge is $(X_{i/t}+1)-X_{i/t}\neq ln(X_{i/t}+1)-lnX_{i/t}$. In fact, the gap or difference of the former remains constant while that of the latter diminishes as X_t increases. As such, this practice, in earnest, changes the distribution of $X_{i/t}$, as examined in the study. Implying that the intended causal effect of X_t is no longer estimated but that of a different variable altogether, $ln(X_{i/t}+1)$. If the undefined $ln(X_{i/t}=0)$ is the problem, as many researchers claim, a traditional and robust approach is to remove i/t for which $X_{i/t}=0$ from the sample. This would mean removing the entire data points of an entity in a panel dataset, X_{it}. As long as the sample remains large, the true causal effect of $X_{i/t}$ is estimated correctly and the Central Limits Theory (CLT) applies for valid inferences based on the distribution of the estimators. Otherwise, as long as the sample is (as if) random and large $(n\to\infty)$, different samples (and sizes) can estimate the true causal effects under the causal effect assumptions. So, it really does not matter which data point, unit or entity is removed from the sample when the sample is (as if) random and large. The true causal effects can be estimated using a non-zero subsample to avoid taking the natural log of zero. Thus, adding one and taking a log is not necessary. Bellemare and Wichman (2019) have initiated a discussion on this.

This study only concentrates on the use of log returns against other approaches for computing asset returns. This is because the log return is a log difference approximation and this study shows and discusses the consequences of log difference approximation, howbeit, in empirical research. However, by no means is this study claiming that the general use of log approximations is suicidal and problematic. Logarithms, in itself, is a powerful tool that has simplified working with numbers, especially large numbers and reduces data variances (Hao, Peng, & He, 2023; He, Guo, & Yue, 2024; Li, Nie, Ruan, & Shen, 2024; Zhang, Su, Sun, Zhang, & Shen, 2015). The use of logarithmic returns in derivative

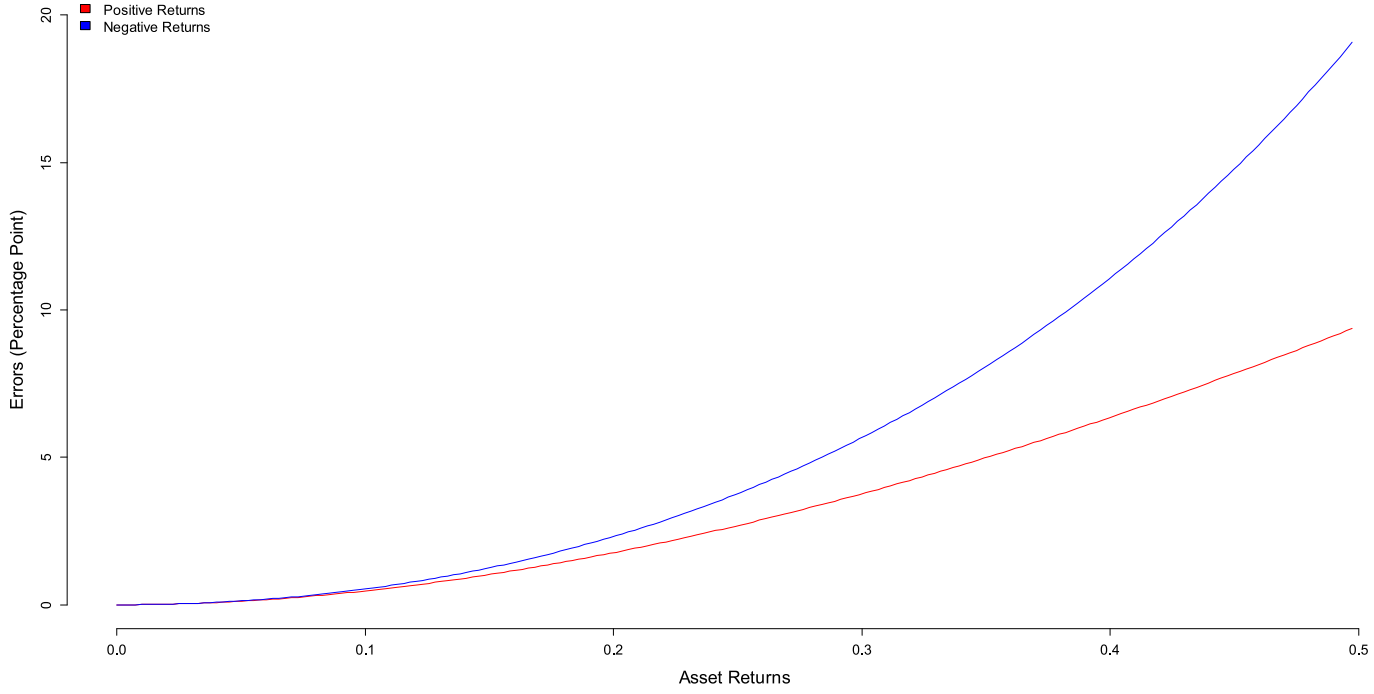

Fig. 1. log return errors.

asset pricing of options and futures is cool and appropriate. This is because the actual values or magnitudes or payoffs of the assets are not (or are less) relevant relative to their relative values or magnitudes or payoffs. Secondly, even when the actual magnitude or payoff of an asset matters, the actual return and log return consistently move in the same direction, thereby giving the same sign despite the possible discrepancies in magnitude or payoff. Beyond these, the logarithmic function is used for model linearization, capturing several nonlinear functional form model relations, estimating elasticities, reducing variances, etc. (Li, Nie, et al., 2024; Tian, Li, & Cheng, 2022). However, the issues discussed in this study become a salient concern when the actual magnitudes or payoffs are relatively important or relevant. As such, caution is required to not invalidate and distort valid inferences.

In theory and practice, the log return approximation is never equal to return. It only approximates returns when the returns tend to zero. Also, the prediction or approximation error of log return increases as returns move away from zero. The log return always under-predicts the actual return of an asset. As such, the log difference is only a conditional approximation of return. In contribution, this article draws researchers' attention to the possible errors of using log approximations in empirical data regression analyses. Particularly, the use of price log differencing to approximate asset returns (log return), the use of log difference on control variables to measure growth rate or percentage change, and the common practice of adding one to a variable before taking natural log to avoid the log of zero. These illustrations buttress the issues, consequences and conditional use of log difference approximations. Nevertheless, this study shows that these issues are not limited to the conditional strong or weak approximations of return or on other (control) variables but include the biasedness and inefficiency of OLS estimators when the (necessary and sufficient) conditions of using log difference approximation are not satisfied. Suffice it to say that a zero mean return is only a necessary but not sufficient condition for a log difference approximation. A sufficient condition is that the variance (or standard deviation) of the return also tends to zero. These are illustrated using the Monte Carlo simulation and empirical data results in Section 3.

2. Data and model

2.1. The log difference approximation

$$r_t = g_t = \frac{P_t - P_{t-1}}{P_{t-1}} \tag{1}$$

The return of an asset is by definition, the change in the current value of an asset relative to its past value. The value of an asset is often captured by its price. As such, return is typically a percentage change or growth rate. Therefore, Eq. (1) captures the actual or true return while Eq. (2) is the log difference approximation of Eq. (1).

$$F(g_t) = lnP_t - lnP_{t-1} = ln\left(1 + \frac{P_t - P_{t-1}}{P_{t-1}}\right) = ln(1 + g_t) = x_t \tag{2}$$

The question becomes, is x_t equal to r_t or directly put, is $ln\left(1 + \frac{P_t - P_{t-1}}{P_{t-1}}\right)$ equal to $\frac{P_t - P_{t-1}}{P_{t-1}}$? Generally, these two are not equal but $ln\left(1 + \frac{P_t - P_{t-1}}{P_{t-1}}\right)$ and approximate $\frac{P_t - P_{t-1}}{P_{t-1}}$ only when $\frac{P_t - P_{t-1}}{P_{t-1}} \to 0$ nonetheless. That is, the log difference approximation of return only works well (i.e. becomes a good approximation) when the return itself is very close to zero, otherwise, the log difference approximation is bad and should be avoided. To appreciate and romance this approximation condition, let's employ the Taylor series proximation for eq. (2), up to the 6th approximation, as shown in eq. (3). In eq. (3) the log difference function, $F(g_t)$, is approximated at a constant a. These Taylor approximations of the log difference function are illustrated in Fig. 2 and Fig. 3 for $g \in [-2, 2]$ and $a \in [-0.99999, 2]$.

$$\begin{aligned} F(g) =& ln(1+a) + \left(\frac{g-a}{1+a}\right) - \frac{1}{2}\left(\frac{g-a}{1+a}\right)^2 + \frac{1}{3}\left(\frac{g-a}{1+a}\right)^3 - \frac{1}{4}\left(\frac{g-a}{1+a}\right)^4 \\ &+ \frac{1}{5}\left(\frac{g-a}{1+a}\right)^5 - \frac{1}{6}\left(\frac{g-a}{1+a}\right)^6 + \dots \end{aligned} \tag{3}$$

Fig. 2 shows the actual values of g, with log difference approximations or predictions of g, with the first and second Taylor approximation predictions of the log difference function. Fig. 3 shows the rest of the

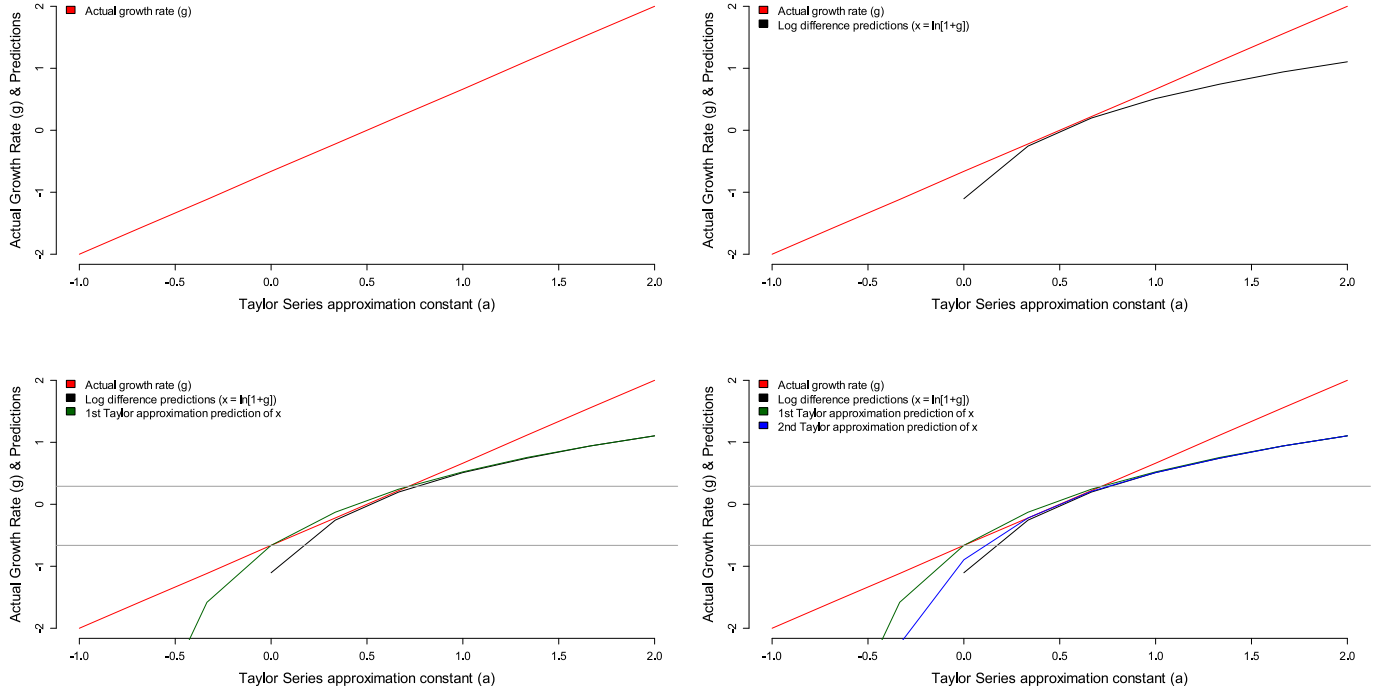

Fig. 2. Returns and og Difference.

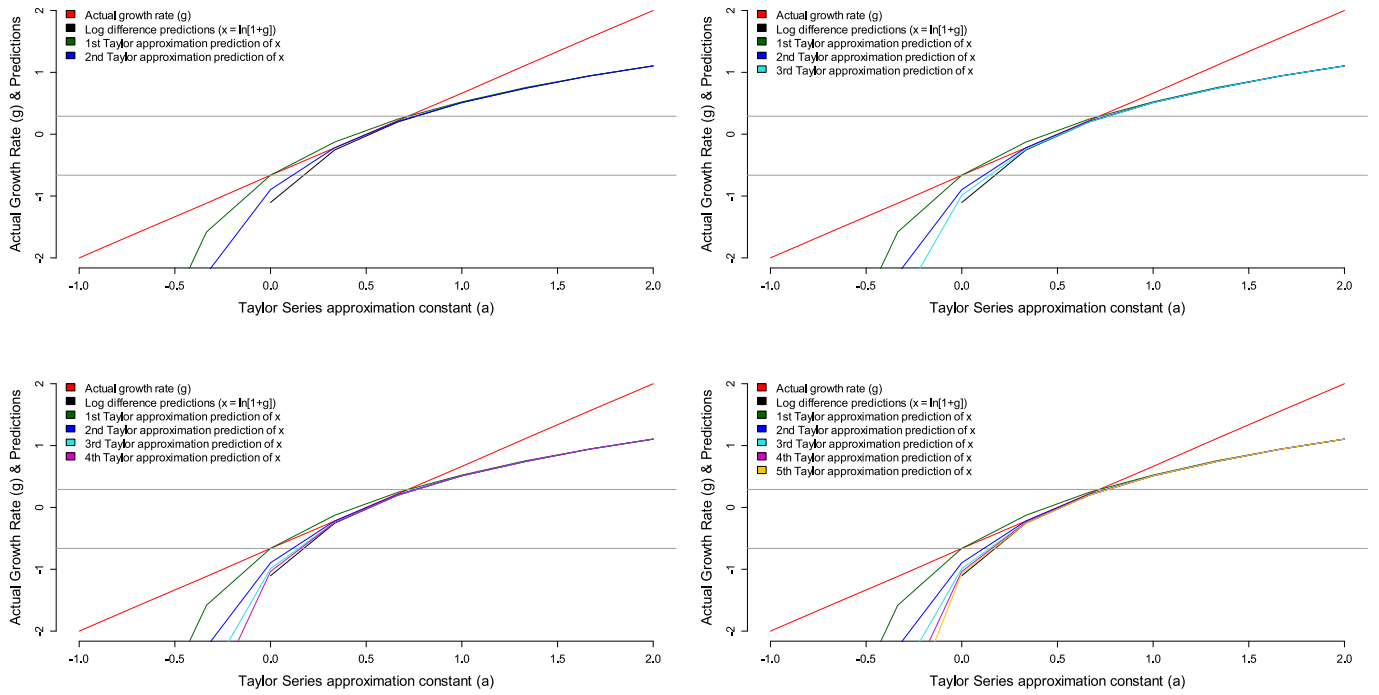

Fig. 3. Returns, Log Difference, and Taylor approximations.

Taylor approximation predictions up to the fifth approximation. There are a few outstanding facts from Figs. 1 and 2. These include that the log difference approximations, x, are never equal to the actual growth, g, but can be a good approximator only when $g \to 0$. Secondly, there are no log difference approximation predictions when $g \leq -1$. The log difference approximation prediction errors increase (the gap between the red and black lines) as g goes away from zero. Thirdly, log differences always under-predict the actual growth or return of an asset. Conversely, the Taylor series approximations exist for $g \leq -1$ and it is capable of over-predicting the actual return. This overprediction diminishes for higher-order Taylor approximations. These make the log difference a conditional approximator of return. To use log difference, one needs to ensure that each period's return is closer to zero. To guarantee that each period's return is closer to zero, the mean and variance (or standard deviation) of the asset's return should tend to zero. That is, the mean being closer to zero is necessary but the variance should be sufficiently closer to zero too. Nevertheless, the safest approach is to compute the return using Eq. (1).

Given that the approximations and predictions of returns, using the log difference, work better when the return is around zero, to minimize the prediction error, we can take the limit of eq. (3) as a tends to 0, $a \to 0$, to derive eq. (4). When the mean and variance of an asset's return are not sufficiently close to zero, log difference produces biased, less precise and less accurate (i.e., less efficient) OLS estimates which leads to misleading results, inference, and conclusions. This is discussed in detail in the subsequent section.

$$\lim_{a \to 0} F(g) = g - \frac{1}{2}g^2 + \frac{1}{3}g^3 - \frac{1}{4}g^4 + \frac{1}{5}g^5 - \frac{1}{6}g^6 + \ldots \tag{4}$$

2.2. How small is small?

The necessary and sufficient conditions to use the log difference approximation, like the log return, is that the mean and variances are small (tend to zero) respectively. Thus, the question becomes how small should the mean and variance of an asset's return be before the log return approximation can yield an unbiased and consistent causal estimate? Based on the simulation results in section 3.1, a mean and variance of at most $\mu \leq |0.1\%|$ and $\sigma^2 \leq 0.1\%^2$ respectively can produce unbiased and consistent causal estimates when the log return approximation is used. Moreover, the grid search results in Fig. 4 and Fig. 5 confirm that return and log return causal effect estimates coincide and converge when $\mu = 0$ and $\sigma^2 = 2.6\%^2$ respectively. This convergence property is also true for different sample sizes, n. Therefore, one could test whether to use log return by either of these two pairs of mean and variance tests. However, all the tests for whether to use log difference approximation presented in this study are for the mean and variance pair, $\mu = 0$ and $\sigma^2 \leq 2.6\%^2$. That is, the null hypothesized parameter set is $(\mu_0 = 0, \sigma^2{}_0 \leq 0.026^2)$ for testing the null hypothesis of using the log difference approximation. This implies that log difference approximations like the log return can be used to estimate the true unbiased and consistent causal effect of the interest variable(s) in a regression model when there is no evidence against this null hypothesis, i.e., this null hypothesis is not rejected at a prespecified test size or level of significance, $\propto$.

To further illustrate the convergence in the causal effects of return and log return when $\mu = 0$ and $\sigma^2 = 2.6\%^2$, Table 1 presents the simulation results for return and log return based on these mean and variance values and under homoscedastic error variance. Based on the results in Table 1, one can see the striking similarities between the coefficients and their standard errors for return and log return. These findings are also valid under heteroscedastic error variance. It's important to note that these estimates, both from return and log return, are unbiased and consistent. Implies a rule of thumb to use either the actual return or the log return when their causal effects have no significant statistical difference (i.e., the causal effects are statistically equal or statistically not different from each other), otherwise, use the actual return to estimate the true causal effects.

Data Generating Function (DGF) is $Y_t = 5 + 50.2r_t + e_t$ *and* $e_t \sim N(0, 1^2)$. *The OLS averages & bootstrap standard errors are reported.*

2.3. When can log difference approximations be used in causal regression analysis?

A possible straight response to this question is; when both the mean

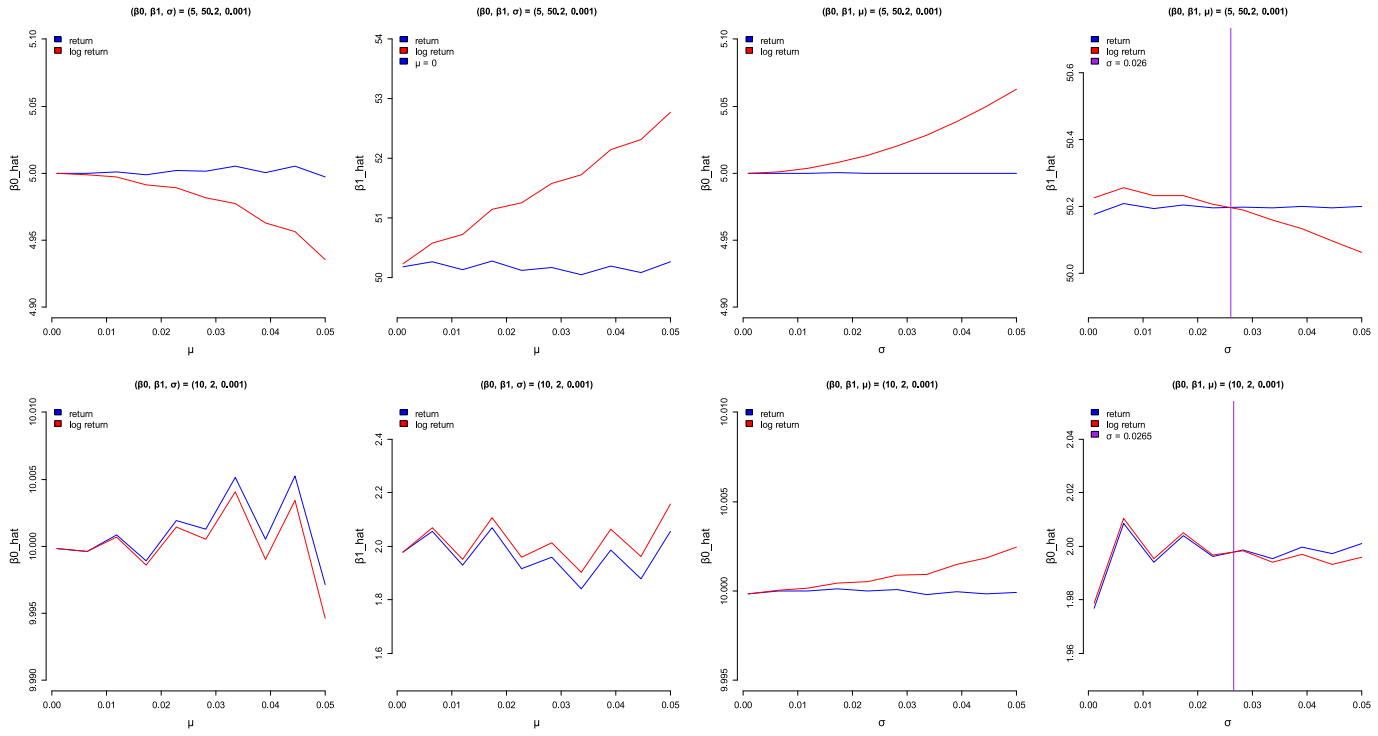

Fig. 4. Grid search for the hypothesized μ *and* σ, $n = 5000$.

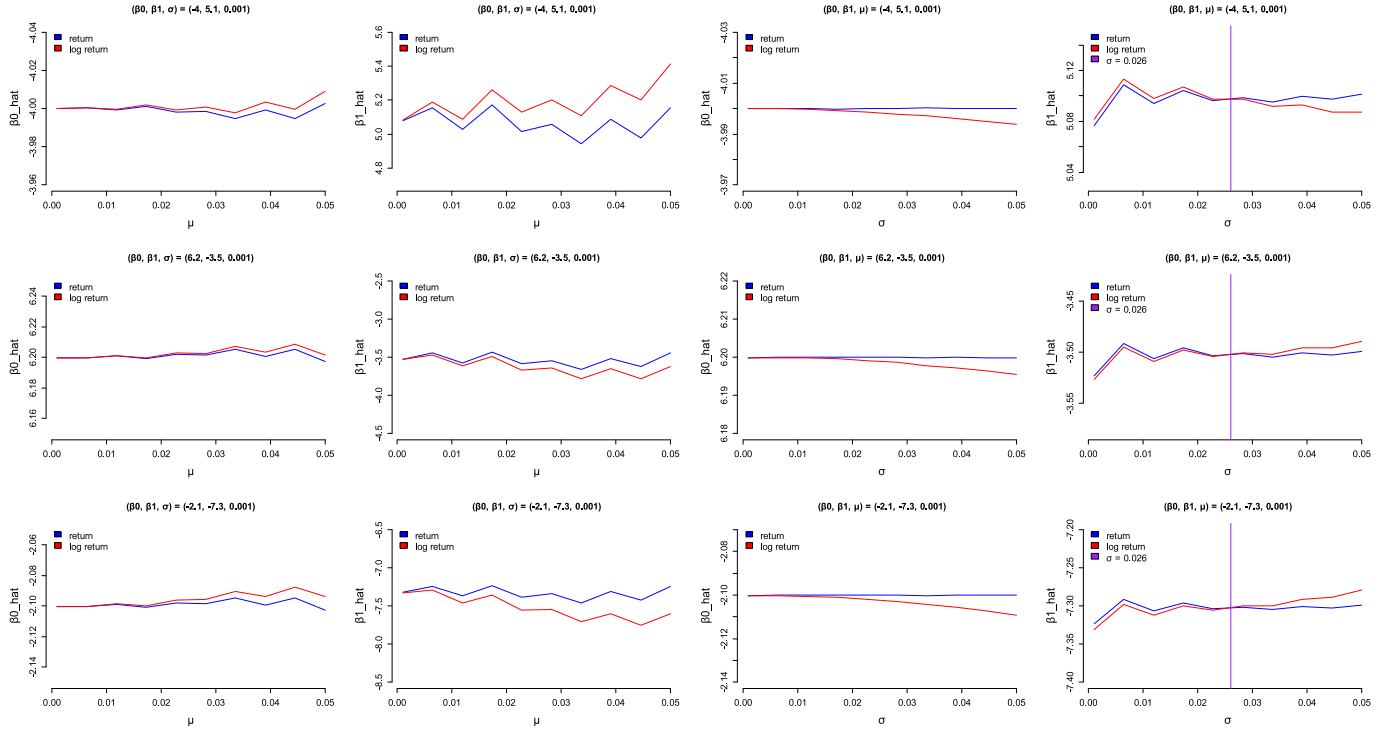

Fig. 5. Another grid search for the hypothesized μ *and* σ, $n = 5000$

and variance of the original series are simultaneously close to zero. This means testing a joint or simultaneous hypothesis on the location (first moment) and scale (centralized second moment) parameters of a univariate sample. A handful of tests have been proposed for this simultaneous test for one sample (Arnold & Shavelle, 1998; Chen & Gao, 2011; Choudhari, Kundu, & Misra, 2001; Park, 2015; Pesarin & Salmaso, 2010; Rao, 1973) and two samples (Duran, Tsai, & Lewis, 1976; Lepage, 1971; Lepage, 1973; Neuhäuser, Leuchs, & Ball, 2011; Rublik, 2009). For a random (stationary) sample, $X_1, X_2, X_3, \ldots, X_n$, drawn from a homoscedastic and normally distributed population with mean, μ, and variance, σ^2, from the parameter space defined as $\boldsymbol{\theta} = \left\{\theta = \left(\mu, \sigma^2\right) : \mu \, \varepsilon \, \mathscr{R}, \sigma^2 \, \varepsilon \, \mathscr{R}^+\right\}$, and we are interested in testing the simultaneous or union and intersection null and alternative hypotheses:

Table 1
$r_t \sim norm(\mu = 0, \sigma = 0.026)$ and $var(e_t|r_t) = \sigma^2$

	$N = 100{,}000$ iterations							
	$n = 5$	$n = 20$	$n = 30$	$n = 50$	$n = 100$	$n = 500$	$n = 5000$	$n = 10000$
Actual Return (r_t, from eq. (1))								
$\widehat{\beta}_0$	4.995473	5.001184	4.999277	4.99717	4.999949	4.999585	5.000067	5.000052
$se(\widehat{\beta}_0)$	0.554667	0.229139	0.187322	0.144339	0.100725	0.045122	0.014112	0.009964
$\widehat{\beta}_1$	50.08972	50.28836	50.09036	50.14718	50.16565	50.21757	50.1984	50.19901
$se(\widehat{\beta}_1)$	27.0433	9.518786	7.361725	5.618159	3.918977	1.75251	0.544311	0.389889
Log Difference (x_t, from eq. (2))								
$\widehat{\beta}_0$	5.005544	5.016367	5.015037	5.013504	5.016588	5.016486	5.017032	5.017015
$se(\widehat{\beta}_0)$	0.554736	0.22919	0.187392	0.144421	0.100789	0.045146	0.014122	0.009966
$\widehat{\beta}_1$	50.06572	50.24591	50.0421	50.09734	50.11737	50.16638	50.14763	50.14788
$se(\widehat{\beta}_1)$	27.03595	9.519329	7.366975	5.621756	3.919652	1.753298	0.544389	0.390152

$$H_0 : (\mu, \sigma^2) = (\mu_0, \sigma_0^2) = \theta_0 \equiv \{\mu = \mu_0\} \cap \{\sigma^2 = {\sigma^2}_0\}$$

$$H_1 : (\mu, \sigma^2)\ \varepsilon\ \boldsymbol{\theta} \neq \theta_0 \equiv \{\mu \neq \mu_0\} \cup \{\sigma^2 \neq \sigma_0^2\}$$

The likelihood ratio testing technique, defined in eq. (5) with density function (Choudhari et al., 2001), Eq. (6), have been shown to test this hypothesis where $0 < x < \infty$ and $a_1(x) < a_2(x)$ are the roots that solve the equation $nlnt - t + x - nln + n = 0$.

$$\lambda(x) = \frac{l(\mu_0, {\sigma^2}_0|x)}{sup_{(\mu,\sigma^2|x)\ \varepsilon\ \theta} l(\mu, \sigma^2|x)} \tag{5}$$

$$g(x) = \frac{1}{2^{0.5n}\sqrt{\pi}\Gamma(0.5(n-1))} e^{-0.5x+0.5n(lnn-1)} \int_{a_1(x)}^{a_2(x)} \frac{1}{v^{1.5}\sqrt{nlnv - v + x - nln + n}} dv \tag{6}$$

The likelihood ratio tests for this simultaneous or union-intersection hypothesis testing generally involve the one-sample mean and variance hypothesis testing statistics. These partial test statistics, eqs. (7) and (8), are not independent (when $n \geq 6$) with a non-zero correlation coefficient, in eq. (9). The test statistic A tests the partial null hypothesis that $H_0 : \mu = \mu_0$ and the fact that the sample variance is an unbiased estimator of the population variance, $E(S^2) = \sigma^2$. Similarly, the test statistic B tests the partial null hypothesis that $H_0 : \sigma^2 = \sigma_0^2$. The correlation coefficient follows directly from the fact that $A^* = \frac{(n-1)(\overline{X}-\mu_0)^2}{\sigma_0^2} \sim {\chi^2}_{\propto,(1)}$ and $E(AB) = (n-1)$. Test statistic A^*, relative to A, requires that the population variance $\sigma^2 = \sigma_0^2$ is known. While the mean and variance of statistic B are its degrees of freedom and twice that respectively, the mean and variance of statistic A are $\frac{n-1}{n-3}$ and $\frac{2(n-1)^2(n-2)}{(n-3)^2(n-5)}$ respectively (Bickel & Doksum, 1977).

$$A = \frac{(n-1)(\overline{X}-\mu_0)^2}{S^2} \sim F_{\propto,(1,n-1)} \overset{d}{\rightarrow} {\chi^2}_{\propto,(1)} \tag{7}$$

$$B = \frac{(n-1)S^2}{\sigma_0^2} \sim {\chi^2}_{\propto,(n-1)} \tag{8}$$

$$corr(A, B) = -\left[\frac{n-5}{n^2-3n+2}\right]^{0.5} \tag{9}$$

Pesarin and Salmaso (2010) show that linear combination strategies like the Liptak, Fisher, and Tipett combination functions can combine these two partially dependent test statistics into a single test for decision-making for or against the null hypothesis. While the Liptak combination function is the unweighted sum of the probability values, π, from these partial hypotheses' tests, the Fisher, Tippett and Liptak combination functions are defined in eq. (10), (11) and (12), where m is the number of partial single restriction dependent tests.

$$\text{Test I} = \sum_{i=1}^{m} -2ln\pi_i \sim {\chi^2}_{\propto,(2m)} \tag{10}$$

$$\text{Test II} = min(\pi_1, \ldots, \pi_m)\ with\ F(y) = 1 - (1-y)^2, 0 < y < 1 \tag{11}$$

$$\text{Test III} = \sum_{i=1}^{m} \pi_i\ with\ F(y) = 0.5y^2, 0 < y \leq 1 \& 1 - 0.5(2-y)^2, 1 < y \leq 2 \tag{12}$$

Several dependent single restriction hypothesis tests produce higher family error rate than the predetermined test size, $1 - (1-\propto)^m \geq \propto$. One prominent solution to adjust for this family error rate and reduce it to the allowed level of significance is the Bonferroni adjusted or corrected techniques. Other combinations of test statistics A and B exits. For instance, Arnold and Shavelle (1998) and Rao (1973) propose using Test IV in eq. (13), Test V, and eq. (14). Test V follows directly from Test IV and the fact that the sample variance is a consistent estimator of the population variance, that is, the sample variance converges to the population variance in large enough samples, ${S^2} \overset{p}{\rightarrow} \sigma^2$. Finally, Test VI in eq. (15) follows from Arnold and Shavelle (1998).

$$Test\ IV = \frac{n(\overline{X}-\mu_0)^2}{\sigma_0^2} + \frac{n\left(S^2-\sigma_0^2\right)^2}{2\sigma_0^4} \sim {\chi^2}_{\propto,(m)} \tag{13}$$

$$Test\ V = \frac{n(\overline{X}-\mu_0)^2}{S^2} + \frac{n\left(S^2-\sigma_0^2\right)^2}{2S^4} \sim {\chi^2}_{\propto,(m)} \tag{14}$$

$$Test\ VI = n\left[ln\left(\frac{\sigma_0^2}{S^2}\right) + \frac{S^2}{\sigma_0^2}\ \frac{(\overline{X}-\mu_0)^2}{\sigma_0^2} - 1\right] \sim {\chi^2}_{\propto,(m)} \tag{15}$$

Based on eqs. (7), (8), (13), and (15), we can test the log approximation mean condition at its zero limit but cannot test its variance

condition at the zero limit. Based on the grid search for different combination values of $\theta_0 = (\mu_0, \sigma_0^2)$ in section 2.2., these test statistics are used for inferences under the null of using log difference approximation, $H_0 : (\mu, \sigma^2) \leq \theta_0 = (0, 0.026^2) \equiv \{\mu = 0\} \cap \{\sigma^2 \leq 0.026^2\}$, against the alternative hypothesis of not using the log difference approximation, $H_1 : (\mu, \sigma^2)\, \varepsilon\, \boldsymbol{\theta} > (0, 0.026^2) \equiv \{\mu \neq 0\} \cup \{\sigma^2 > 0.026^2\}$. However, it is important to state that since the conditions for log difference approximation is that both mean and variance tend to zero. A partial mean test hypothesized value that tends to zero, in absolute terms, other than the zero limit itself could be applied. This also applies to a partial hypothesized standard deviation limit value that is around 2.6 %.

$H_0 : \beta_r = \beta_x \equiv$ *Use log difference approximation*

$H_1 : \beta_r \neq \beta_x \equiv$ *Do not use log difference approximation*

$$T = \frac{\widehat{\beta}_r - \widehat{\beta}_x}{\sqrt{se^2(\widehat{\beta}_r) + se^2(\widehat{\beta}_x)}} \sim t_{n_r+n_x-2,\alpha/2} \tag{16}$$

An alternative ex-post estimation testing approach is also discussed. Based on the simulation exercise, the actual returns have consistently estimated the true causal effect under the causal assumptions. As such, can be used as a reference for the log return's causal effect estimation. Therefore, a difference in estimates test can be done to decide when to use the log difference approximations, like log return, in a regression analysis. Under the null hypothesis, the causal estimate of return is statistically not different from that of the log difference approximation, log return. Also, under the null, the return and log return estimators are unbiased and consistent, but the return estimator is efficient. Under the alternative hypothesis, only the return estimator is unbiased and consistent. This test is the basic single restriction $T - test$ statistic in eq. (16) on a regression models' estimates. $\widehat{\beta}_r$ is the estimate using the return while $\widehat{\beta}_x$ is the log difference approximation or log return estimate. Secondly, this test statistic in eq. (16) also applies to testing the null hypothesis of using adding one before taking the natural logarithm of a variable, $ln(X+1)$, against the alternative hypothesis of using a non-zero large-enough subsample, in a causal effect regression analysis.

Table 2 presents and discusses the rejection rates or type I errors of the simultaneous mean and variance test statistics. This Type I error, the probability of rejecting a true null hypothesis, in Table 2 is for a 5 % test size or significance level. Therefore, the results in Table 2 are expected to be around 5 %. Return processes from $\mu = 0$ and $\sigma^2 = 0.026^2$ normal population are simulated for different sample sizes and the test statistics I – VI are applied to test the true null hypothesis of using log difference approximations against the false alternative of not using the log difference approximations in $N = 100,000$ iterations. These type I errors in Table 2 are not consistently around 5 % but appear to increase in sample sizes. Consequently, this leads to high test power for the test statistics, subtracting type II errors from one. This is expected since type I error is directly related to the power of the test but inversely related to type II error. Table 2 results confirm that these test statistics are not efficient or sufficient to determine whether (or not) to use the log difference approximation at a predetermined test size or significance level. This creates room for an improved test statistic that can sufficiently test the null hypothesis of using the log difference approximation. However, the proposed alternative difference-in-estimates test in eq. (16) performed very well with a type I error of zero for all sample sizes. As such, the proposed test in eq. (16) is proposed as the benchmark test to test the null hypothesis of using log difference approximations against its alternative hypothesis of not using the log difference approximations.

2.4. Empirical data

To empirically illustrate the discussions of this study on log difference approximations, the daily Bitcoin price and trade volume data, from 1/7/2019 to 30/6/2020, is collected from the coin market capitalization database. This dataset is used to investigate the empirical use of log difference approximations, log returns, in a regression analysis. Secondly, to empirically illustrate the causal estimation effects of $ln(X_{i/t}+1)$, 2019 firm-level empirical data is sourced from CSMAR for a sample of A-share listed companies in China.

3. Results and discussions

3.1. Simulation results and discussions

3.1.1. Returns and log differences under homoscedastic errors

In this section, the performance of the log difference approximation for return, under different scenarios and conditions, is examined using simulation exercises. Generally, the scenario designs are homoscedasticity, heteroscedasticity, and omitting control variable(s). The conditions are small sample mean and variance, small mean and large variance, and large mean and variance. Table 3 shows the baseline simulation results for returns and its log difference approximation under homoscedastic errors. The simulation condition is that of small mean and variance. That is, the returns are normally simulated from an equal population mean and variance of 0.1 % for each of the 100,000 iterations.

Based on Table 3 results, the OLS estimates of the true population slope coefficient while using the actual return, r_t from eq. (1), and the log difference, x_t from eq. (2), are unbiased and consistent. However, the log difference estimates are less efficient. As such, the estimates from using the actual returns, r_t, are BLUE relative to the use of log difference approximation, x_t. The results in Table 3 confirm that when the necessary and sufficient conditions to approximate assets' returns with log difference are met, the slope estimates of a log difference approximation remain unbiased and consistent but less efficient. Hence, the necessary condition to approximate returns with log difference is that the sample mean of the return tends to zero while the sufficient condition is that the variance of the sample returns also tends to zero. Table 3 confirms these log difference approximation conditions.

In Table 4, the returns are again generated from a normal distribution with a 0.1 % population mean but with a relatively larger variance of 2.3. The simulation results in Table 4 reveal that using the actual return, r_t from eq. (1), produced unbiased, consistent and efficient estimates of the true population slope parameter. Conversely, the log difference estimates give both biased and inconsistent estimates of the true slope and intercept terms. Therefore, the OLS estimates remain

Table 2
Tests' type I error rates

	$N = 100,000$ *iterations*							
	$n = 5$	$n = 20$	$n = 30$	$n = 50$	$n = 100$	$n = 500$	$n = 5000$	$n = 10000$
Test I	0.02176	0.03545	0.03762	0.04238	0.04597	0.05785	0.12996	0.21233
Test II	0.04548	0.04923	0.04911	0.05109	0.05086	0.05775	0.12370	0.20141
Test III	0.00407	0.02672	0.03056	0.03625	0.04113	0.05516	0.11214	0.16370
Test IV	0.07053	0.05624	0.05313	0.05339	0.05252	0.05793	0.12289	0.20029
Test V	0.23879	0.11102	0.0914	0.07761	0.06370	0.05777	0.12117	0.1989
Test VI	0.08225	0.05717	0.05435	0.05365	0.05250	0.05725	0.12237	0.20001

Table 3
$r_t \sim norm(\mu = \sigma = 0.001)$ and $var(e_t|r_t) = \sigma^2$

	N = 100,000 iterations							
	$n = 5$	$n = 20$	$n = 30$	$n = 50$	$n = 100$	$n = 500$	$n = 5000$	$n = 10000$
Actual Return (r_t, from eq. (1))								
$\widehat{\beta}_0$	5.000453	4.998705	5.000279	5.000971	4.999866	5.000093	5.000056	4.999965
$se(\widehat{\beta}_0)$	0.898076	0.333491	0.267912	0.204023	0.143103	0.063358	0.020014	0.014115
$\widehat{\beta}_1$	51.34340	50.50396	49.35558	49.82161	50.06522	50.23295	50.18034	50.22028
$se(\widehat{\beta}_1)$	705.9825	242.0610	192.3942	146.2836	101.6518	44.86654	14.111215	9.959523
Log Difference (x_t, from eq. (2))								
$\widehat{\beta}_0$	5.000445	4.998702	5.000277	5.000970	4.999865	5.000093	5.000056	4.999965
$se(\widehat{\beta}_0)$	0.898418	0.333516	0.267925	0.204029	0.143105	0.063358	0.020013	0.014115
$\widehat{\beta}_1$	51.39232	50.55473	49.40492	49.87120	50.11521	50.28315	50.23047	50.27043
$se(\widehat{\beta}_1)$	706.6922	242.3021	192.5860	146.4300	101.7534	44.91128	14.12533	9.969448

Data Generating Function (DGF) is $Y_t = 5 + 50.2r_t + e_t$ and $e_t \sim N(0, 1^2)$. The OLS averages & bootstrap standard errors are reported.

Table 4
$r_t \sim norm(\mu = 0.001, \sigma = 2.3)$ and $var(e_t|r_t) = \sigma^2$

	N = 100,000 iterations					
	$n = 20$	$n = 30$	$n = 50$	$n = 100$	$n = 500$	$n = 5000$
	Actual Return (r_t, from eq. (1))					
$\widehat{\beta}_0$	4.999124	4.999763	5.000682	4.999803	5.000119	4.999304
$se(\widehat{\beta}_0)$	0.243316	0.197516	0.151773	0.107316	0.047801	0.015271
$\widehat{\beta}_1$	50.20006	50.19941	50.19988	50.19989	50.20001	50.19946
$se(\widehat{\beta}_1)$	0.157916	0.124237	0.093237	0.064594	0.028409	0.009184
	Log Difference (x_t, from eq. (2))					
$\widehat{\beta}_0$	39.75207	40.26754	40.69823	41.02722	41.30717	41.35817
$se(\widehat{\beta}_0)$	8.437654	6.786831	5.224683	3.679969	1.621912	1.621912
$\widehat{\beta}_1$	62.58866	62.40113	62.08694	61.75775	61.39941	61.32379
$se(\widehat{\beta}_1)$	11.59787	9.948291	8.241095	6.230963	3.030842	0.961507

Data Generating Function (DGF) is $Y_t = 5 + 50.2r_t + e_t$ and $e_t \sim N(0, 1^2)$. The OLS averages & bootstrap standard errors are reported.

BLUE using the actual return, r_t, from eq. (1) while the log difference approximation fails even when the mean return is very close (or tends) to zero, 0.001. This suggests that having the sample average of the actual returns tending to zero is only a necessary condition and not a sufficient condition to adequately approximate the actual returns with log difference, x_t, from eq. (2). Both the necessary and sufficient conditions are illustrated in Table 1. As such, in addition to the sample mean being close to zero, the variance of the sample return should also be sufficiently close to zero. Based on the last column in Table 4, even with a large sample size, the log return estimates do not converge to the population parameter. As such, the OLS estimators for log return are inconsistent.

Given the normally simulated returns with relatively larger variance, in large sample sizes, the computed asset prices become explosive and can tend to either zero or infinity. This is the reason larger sample size results are not reported in Table 4. To work around this issue, the asset prices are directly simulated from a chi-square distribution with mean return restriction or condition. That is, for each iteration, the asset prices are simulated from the chi-square distribution, next, the actual returns are computed following eq. (1). Then, the sample average of these returns is checked against the mean return restriction. If this restriction is satisfied, the OLS estimation is executed, otherwise, another asset

Table 5
$P_t \sim \chi^2{}_5|[0.5 \leq \overline{r_t} \leq 2]$ and $var(e_t|r_t) = \sigma^2$

	N = 100,000 iterations							
	$n = 5$	$n = 20$	$n = 30$	$n = 50$	$n = 100$	$n = 500$	$n = 5000$	$n = 10000$
Actual Return (r_t, from eq. (1))								
$\widehat{\beta}_0$	5.001184	4.999541	4.999490	4.999312	4.999784	5.000068	5.000001	5.000022
$se(\widehat{\beta}_0)$	0.504371	0.238711	0.193154	0.149210	0.104847	0.046405	0.0145947	0.010317
$\widehat{\beta}_1$	50.19984	50.20033	50.19996	50.20026	50.19996	50.19988	50.20000	50.20001
$se(\widehat{\beta}_1)$	0.295004	0.113749	0.090187	0.068317	0.046735	0.019663	0.005454	0.003744
Log Difference (x_t, from eq. (2))								
$\widehat{\beta}_0$	44.53044	46.19288	45.05813	43.25560	41.15670	38.62279	38.46406	38.47185
$se(\widehat{\beta}_0)$	22.23387	15.69305	14.27221	12.34472	9.762046	5.214240	1.779176	1.266048
$\widehat{\beta}_1$	81.43259	89.80402	91.39902	92.05184	91.84194	90.48083	90.89022	90.97349
$se(\widehat{\beta}_1)$	24.08447	27.56551	27.99995	27.47690	25.40035	17.38078	7.236477	5.325233

Data Generating Function (DGF) is $Y_t = 5 + 50.2r_t + e_t$ and $e_t \sim N(0, 1^2)$. The OLS averages & bootstrap standard errors are reported.

price is simulated. This is solely done to be able to control the sample mean of the simulated returns. Generally, Table 1 illustrates the scenario where the sample mean and variance of the returns tend to zero. In Table 4, the sample mean of the return is maintained to tend to zero but the variance is increased. In Table 5, the mean and variance of the sample returns are relatively larger and do not tend to zero. In the simulated samples, the mean of the returns is restricted to lie between 0.5 and 2, with both limits, while on average, the standard deviations of the sampled return are 2.133359, 2.429677, 2.483547, 2.509065, 2.525344, 2.567651, 2.748710, and 2.790675 for each of the sample sizes respectively. In essence, Table 5 illustrates the condition of a relatively larger sample return mean and variance. As shown in Table 5, the actual returns results are unbiased and consistent while that of the log difference is biased and inconsistent even in large samples. Based on the last column in Table 5, even with a large sample size, the log return estimates do not converge to the population parameter. As such, the OLS estimators for log return are inconsistent.

3.1.2. *Returns and log differences under heteroscedastic errors*

Under this second scenario, all the conditions simulated in Tables 1 to 3, are repeated but with heteroscedastic errors. A simple heteroscedastic nature of the error term is applied. First, the simulated returns are grouped by quantiles (0 %, 25 %, 50 %, 75 %, and 100 %). Then, the error is simulated with zero mean and a variance that increases in the quantiles. Specifically, if the return is between the 0 % to 25 % quantile, the variance of the simulated error is 1, if the return is between the 25 % to 50 % quantile, the variance of the simulated error is 5, if the return is between the 50 % to 75 % quantile, the variance of the simulated error is 10, and if the return is between the 75 % to 100 % quantile, the variance of the simulated error is 18. The results in Table 6 are the heteroscedasticity version of the homoscedasticity results in Table 1. As expected, the OLS estimates remain unbiased and consistent in the presence of heteroscedasticity. That is, the actual return estimates and that of the log difference are both unbiased and consistent in large samples. This is due to the fact the necessary and sufficient conditions for log difference approximation of return are satisfied. However, relative to the log difference approximation results, the actual return estimates of the slope are more precise and more accurate with more efficiency.

The results in Table 7 are the heteroscedasticity version of the homoscedasticity results in Table 4. From the results in Table 7, the estimates with the actual returns are unbiased and consistent given heteroscedastic errors while that of the log difference are biased and inconsistent. That is, even when only the necessary conditions for approximating returns with log difference are met, yet, log difference approximation produces biased and inconsistent results thus, leading to invalid inference, misleading conclusions, recommendations and policy implications. Based on the last column in Table 7, even with a large sample size, the log return estimates do not converge to the population parameter. As such, the OLS estimators for log return are inconsistent. Again, this reinforces that whether there is homoscedasticity or heteroscedasticity in the model, as long as the necessary conditions and sufficient conditions of log difference approximation are met, the log difference estimates remain unbiased and consistent (Table 1 and Table 6). However, when the sufficient condition is not satisfied, the log difference approximation results are biased and inconsistent (s 2 and 5), whether or not the error is homoscedastic.

Similarly, Table 8 results are the heteroscedasticity version of the homoscedasticity results in Table 5. In the samples, the mean of the returns lies between 0.5 and 2, with both limits included, while the sample returns standard deviations, on average, are 2.133359, 2.429677, 2.483547, 2.509065, 2.525344, 2.567651, 2.748710, and 2.790675 respectively for each of the simulated sample sizes. The results in Table 8 confirm, yet again, that the log difference estimates are biased and inconsistent, even in large samples, when the necessary and sufficient conditions of approximating returns with log difference are not met, given heteroscedasticity or homoscedasticity. Based on the last column in Table 8, even with a large sample size, the log return estimates do not converge to the population parameter. As such, the OLS estimators for log return are inconsistent.

3.1.3. *Log difference approximations, control variables and* $ln(X_{i/t}+1)$

While omitting some control variables from a regression model can bias the causal effect estimation of the interest regressor or independent variable, omitting other control variables does not bias the causal effect estimation of the interest regressor. If a control variable correlates with the main interest regressor and explains the dependent variable, in part, omitting such a control variable in the regression model can bias the causal effect estimation of the interest regressor. This is called the Omitted Variable Bias (OVB) which leads to endogeneity problems in the regression model. Thereby, biasing the OLS estimation of the true causal effect of the interest regressor in a regression model. On the contrary, if a control variable either correlates with the main interest regressor or explains the dependent variable, in part, omitting such a control variable in the regression model does not bias the causal effect estimation of the interest regressor. Thus, the true unbiased and consistent causal effect can still be estimated using OLS and under the causal assumptions. The first part of this third scenario design is developed to investigate the performance of log difference given these two possible omitted variables in a regression model. The second part answers the question of whether it matters to use the log difference approximation for a control variable. While the last part illustrates the dangers of $ln(X_{i/t}+1)$ and offered a robust solution.

When W_t, $[W_t : \beta_2 \neq 0 \; and \; cov(r_t, W_t) \neq 0]$, is omitted from the regression model, the OLS estimator for both the actual returns and log

Table 6
$r_t \sim norm(\mu = \sigma = 0.001)$ and $var(e_t|r_t) = {\sigma_i}^2$

	N = 100,000 iterations							
	$n = 5$	$n = 20$	$n = 30$	$n = 50$	$n = 100$	$n = 500$	$n = 5000$	$n = 10000$
Actual Return (r_t, from eq. (1))								
$\widehat{\beta}_0$	5.023936	4.998142	4.998252	4.997849	4.998206	5.000470	5.000305	4.999794
$se(\widehat{\beta}_0)$	6.846201	1.876920	1.456227	1.068415	0.735633	0.320239	0.100862	0.071052
$\widehat{\beta}_1$	43.43653	44.82772	40.80869	54.51218	48.60060	51.03428	50.00387	50.29240
$se(\widehat{\beta}_1)$	8364.014	2996.166	2405.580	1818.003	1260.744	559.1872	175.7574	124.5387
Log Difference (x_t, from eq. (2))								
$\widehat{\beta}_0$	5.023944	4.998130	4.998245	4.997848	4.998206	5.000470	5.000305	4.999794
$se(\widehat{\beta}_0)$	6.850775	1.876267	1.455442	1.067650	0.735007	0.319924	0.100761	0.070981
$\widehat{\beta}_1$	43.46754	44.87742	40.85409	54.56581	48.64880	51.08545	50.05417	50.34241
$se(\widehat{\beta}_1)$	8371.696	2998.181	2407.121	1819.095	1261.457	559.4873	175.8513	124.6050

Data Generating Function (DGF) is $Y_t = 5 + 50.2r_t + e_t$ and $e_t \sim N(0, {\sigma_i}^2)$. The OLS averages & bootstrap standard errors are reported.

Table 7
$r_t \sim norm(\mu = 0.001, \sigma = 2.3)$ and $var(e_t|r_t) = {\sigma_i}^2$

	N = 100,000 iterations					
	$n = 20$	$n = 30$	$n = 50$	$n = 100$	$n = 500$	$n = 5000$
			Actual Return (r_t, from eq. (1))			
$\widehat{\beta}_0$	4.996886	4.994354	5.001707	4.995691	5.001042	4.988565
$se(\widehat{\beta}_0)$	1.667356	1.366501	1.032978	0.717815	0.317345	0.099705
$\widehat{\beta}_1$	50.19579	50.19209	50.20253	50.19984	50.20044	50.19035
$se(\widehat{\beta}_1)$	2.298701	1.824227	1.373061	0.950353	0.420960	0.135216
			Log Difference (x_t, from eq. (2))			
$\widehat{\beta}_0$	39.74612	40.25663	40.70065	41.02289	41.30848	41.34070
$se(\widehat{\beta}_0)$	8.837060	7.117623	5.476249	3.851575	1.698864	0.555889
$\widehat{\beta}_1$	62.58124	62.39144	62.08946	61.75602	61.40023	61.31032
$se(\widehat{\beta}_1)$	11.84360	10.14540	8.376539	6.311155	3.064428	0.969547

Data Generating Function (DGF) is $Y_t = 5 + 50.2r_t + e_t$ and $e_t \sim N(0, {\sigma_i}^2)$. The OLS averages & bootstrap standard errors are reported.

Table 8
$P_t \sim {\chi^2}_5 | [0.5 \leq \overline{r_t} \leq 2]$ and $var(e_t|r_t) = {\sigma_i}^2$

	N = 100,000 iterations							
	$n = 5$	$n = 20$	$n = 30$	$n = 50$	$n = 100$	$n = 500$	$n = 5000$	$n = 10000$
Actual Return (r_t, from eq. (1))								
$\widehat{\beta}_0$	5.010682	4.999215	4.996834	4.993466	4.998089	5.000201	4.999979	5.000202
$se(\widehat{\beta}_0)$	2.275419	1.905542	1.624339	1.279679	0.917854	0.420502	0.136630	0.097356
$\widehat{\beta}_1$	50.19430	50.20475	50.19740	50.20343	50.19884	50.19820	50.19995	50.20014
$se(\widehat{\beta}_1)$	3.987202	1.783827	1.442266	1.103901	0.767636	0.328276	0.092800	0.063939
Log Difference (x_t, from eq. (2))								
$\widehat{\beta}_0$	44.53494	46.19534	45.05364	43.25109	41.15471	38.62181	38.46401	38.47212
$se(\widehat{\beta}_0)$	22.56433	15.86815	14.39862	12.42838	9.823160	5.239167	1.785523	1.270212
$\widehat{\beta}_1$	81.42533	89.80783	91.39343	92.05231	91.83958	90.47835	90.89025	90.97367
$se(\widehat{\beta}_1)$	24.70016	27.69189	28.08448	27.52214	25.43510	17.39403	7.238879	5.326243

Data Generating Function (DGF) is $Y_t = 5 + 50.2r_t + e_t$ and $e_t \sim N(0, {\sigma_i}^2)$. The OLS averages & bootstrap standard errors are reported.

Table 9
$r_t \sim norm(\mu = \sigma = 0.001)$, $var(e_t|r_t) = \sigma^2$, and controls

	N = 100,000 iterations							
	$n = 5$	$n = 20$	$n = 30$	$n = 50$	$n = 100$	$n = 500$	$n = 5000$	$n = 10000$
Actual Return (r_t, from eq. (1))								
$\widehat{\beta}_0$	3.819122	3.821517	3.821660	3.820868	3.821430	3.820866	3.820977	3.820937
$se(\widehat{\beta}_0)$	1.144578	0.423638	0.340861	0.259879	0.181317	0.080477	0.025524	0.017983
$\widehat{\beta}_1$	50.57962	50.05730	49.55907	50.14266	50.09642	50.04591	50.31307	50.23437
$se(\widehat{\beta}_1)$	896.9574	307.3151	245.2884	186.1672	128.8637	56.94253	18.03241	12.72858
Log Difference (x_t, from eq. (2))								
$\widehat{\beta}_0$	3.819112	3.821516	3.821659	3.820867	3.821430	3.820866	3.820977	3.820937
$se(\widehat{\beta}_0)$	1.145022	0.423670	0.340877	0.259887	0.181320	0.080477	0.025524	0.017983
$\widehat{\beta}_1$	50.63068	50.10650	49.60877	50.19273	50.14623	50.09608	50.36330	50.28456
$se(\widehat{\beta}_1)$	897.8652	307.6206	245.5337	186.3530	128.9926	56.99942	18.05048	12.74133

Data Generating Function (DGF) is $Y_t = 5 + 50.2r_t - 0.5W_t + e_t$, $cov(r_t, W_t) = 0$, and $e_t \sim N(0, 1^2)$. The OLS averages & bootstrap standard errors are reported.

return are biased and inconsistent. As such, whether the actual returns or the log difference approximation is used, the OLS estimators remain biased, even in large samples, in the presence of OVB. This is as expected since OVB leads to an endogeneity problem in the regression model. Thus, the simulation results are not presented. Conversely, omitting W_t, $[W_t : \beta_2 \neq 0 \; and \; cov(r_t, W_t) = 0]$ and $[W_t : \beta_2 = 0 \; and \; cov(r_t, W_t) \neq 0]$ from the regression model reveals that the log difference approximation gives an unbiased, consistent, and inefficient OLS estimator when the necessary and sufficient conditions for the log difference approximation of returns are met (see Table 9). On the other hand, the log difference approximation produces biased estimates when the necessary and sufficient conditions of approximating returns with log differences are not met (see Tables 8 and 9).

From Table 9 and as expected, since the control variable, W_t, has no relationship with the interest regressor, r_t, omitting the covariate, W_t, in the sample regression model will not bias the slope estimate, but the intercept term. The simulation results in Table 9 reveal that this is true for both the actual return and the log difference approximation due to the fact the necessary and sufficient conditions for the log difference approximations are met. However, the log difference estimates remain

Table 10
$r_t \sim norm(\mu = 0.001, \sigma = 2.3)$, $var(e_t|r_t) = \sigma^2$, and controls

	$N = 100{,}000$ iterations				
	$n = 20$	$n = 30$	$n = 50$	$n = 100$	$n = 500$
Actual Return (r_t, from eq. (1))					
$\widehat{\beta}_0$	3.821238	3.821133	3.820899	3.821385	3.820794
$se(\widehat{\beta}_0)$	0.309367	0.251063	0.193284	0.136247	0.060799
$\widehat{\beta}_1$	50.20005	50.19962	50.19987	50.19998	50.19984
$se(\widehat{\beta}_1)$	0.201036	0.157897	0.118974	0.081789	0.036009
Log Difference (x_t, from eq. (2))					
$\widehat{\beta}_0$	38.53796	39.13200	39.54394	39.87877	40.12964
$se(\widehat{\beta}_0)$	8.423840	6.831013	5.236542	3.672762	1.62646
$\widehat{\beta}_1$	62.52280	62.39477	62.06056	61.70622	61.39991
$se(\widehat{\beta}_1)$	11.59268	9.980341	8.221649	6.265124	3.027167

Data Generating Function (DGF) is $Y_t = 5 + 50.2r_t - 0.5W_t + e_t$, $cov(r_t, W_t) = 0$, and $e_t \sim N(0, 1^2)$. The OLS averages & bootstrap standard errors are reported.

less accurate, less precise, and inefficient, even in large samples, relative to the return estimates although both are unbiased and consistent.

Again, the results in Table 10 show that the log difference approximation of return produces a biased and inconsistent estimate of the true slope estimate when the control variable, W_t, which has no relationship with the interest variable, is omitted. However, this is largely due to the fact the necessary and sufficient conditions for using the log difference approximation of return are not jointly satisfied.

In the same light, with relatively larger sample return mean and variance, Table 11 results confirm that the return slope estimates remain unbiased and consistent while that of the log difference is biased and inconsistent even in large samples. The simulated return restriction is that the sample average lies between 0.5 and 2, with both limits included while on average, the standard deviations for each sample size are 2.132211, 2.428752, 2.478786, 2.509340, 2.528713, 2.572744, 2.755755, and 2.790350 respectively. Based on the last column in Table 11, even with a large sample size, the log return estimates do not converge to the population parameter. As such, the OLS estimators for log return are inconsistent.

The use of log difference approximations (like the log return) as a regressor is still problematic, whether (or not) it is the main interest regressor. When the log difference approximation conditions are not met, the use of the log difference approximation to compute the main interest regressor(s) leads to a biased and inconsistent causal OLS estimator. Also, when the log difference approximation conditions are (not) met, the use of log difference approximation to compute other covariates, control variables, in the regression model can lead to biased and inconsistent causal OLS estimator of the interest regressor(s) if this log differenced control variable correlates with the interest regressor and (significantly) explains the dependent variable, in part. This is the OVB that leads to an endogeneity problem in the regression model. However, if this log-differenced covariate is a control variable that either correlates with the interest regressor or (significantly) explains the dependent variable, in part, the OLS causal estimation of the interest variable remains unbiased and consistent, under the causal assumptions, when the log difference approximation conditions are unmet.

Therefore, it's important to use the right log difference approximation (i.e., meeting the conditions) for control variables, especially when the log differenced control variable correlates with the interest regressor (s) and explains the dependent variable, in part. In a nutshell, if a covariate in a regression model correlates with the main interest regressor (s) and (significantly) affects or explains the dependent variable, then, it can affect the coefficient of the interest regressor when omitted in the regression model. Therefore, to estimate the true causal effect of the interest regressor, we must include this covariate in the regression model. If we must include this covariate in the regression model, then, it's important the variable is calculated or measured or approximated correctly. Hence, the log differenced covariate needs to be correct, especially, when it correlates with the interest regressor and (significantly) explains the dependent variable, in part. This is because it must be included in the model to correctly estimate the true causal effect of the interest regressor. In fact, the results in Table 12 show that when the log approximation conditions are unmet, the use of log difference approximation as a control variable can, indeed, bias the OLS causal effect estimation of the main interest regressor when they are (significantly) correlated, and the log differenced control (significantly) explains the dependent variable. This is the case where the log differenced control variable, r_t, $[r_t : \beta_2 \neq 0 \; and \; cov(r_t, Z_t) \neq 0]$, is included in the regression model. When the actual return, r_t, is used, the true causal effect of X_t is correctly estimated. But when the log returns, x_t, is used (and does not meet the log approximation conditions), the OLS estimator of the true causal effect of X_t is incorrectly estimated. In conclusion, a log-differenced control variable, W_t, can bias the causal effect estimate of an interest regressor. Secondly, the causal effects of both the actual returns and log returns are biased and inconsistent in the presence of OVB. Finally, omitting a control variable that either explains the dependent variable, in part, or correlates with the main interest regressor does not bias the causal effect of the actual returns but biases that of the log return when the log difference approximations are unmet.

Data Generating Function (DGF) is $Y_i = 2 + 3ln(X_i) + e_i$ and $e_i \sim N(0, 1^2)$. The OLS averages & bootstrap standard errors are reported.

At this point, the dangers of using $ln(X_{i/t} + 1)$ in regression analysis are illustrated. In this simulation and for all sample sizes, $X_i \sim \chi^2{}_7$ and rounded to the nearest whole numbers to have a count or integer data. Next, if any data points in X_i is zero, the random sample, X_i, is drawn

Table 11
$P_t \sim \chi^2{}_5 | [0.5 \leq \overline{r_t} \leq 2]$, $var(e_t|r_t) = \sigma^2$, and controls

	$N = 100{,}000$ iterations							
	$n = 5$	$n = 20$	$n = 30$	$n = 50$	$n = 100$	$n = 500$	$n = 5000$	$n = 10000$
Actual Return (r_t, from eq. (1))								
$\widehat{\beta}_0$	3.822755	3.821860	3.820770	3.820487	3.821355	3.821436	3.820975	3.820981
$se(\widehat{\beta}_0)$	0.640875	0.303620	0.245945	0.189120	0.132799	0.058830	0.018585	0.013112
$\widehat{\beta}_1$	50.19783	50.19976	50.20048	50.20000	50.19996	50.19999	50.20003	50.19999
$se(\widehat{\beta}_1)$	0.376853	0.144922	0.114919	0.087556	0.059818	0.024950	0.006937	0.004767
Log Difference (x_t, from eq. (2))								
$\widehat{\beta}_0$	43.32305	44.98827	43.80607	42.08127	39.94933	37.45529	37.29433	37.28683
$se(\widehat{\beta}_0)$	22.25274	15.63188	14.14862	12.28007	9.773476	5.234795	1.799035	1.274851
$\widehat{\beta}_1$	81.37366	89.80789	91.36153	92.02502	91.96111	90.55170	90.94052	90.95876
$se(\widehat{\beta}_1)$	24.14736	27.59034	27.95930	27.44602	25.77400	17.60233	7.495667	5.419247

Data Generating Function (DGF) is $Y_t = 5 + 50.2r_t - 0.5W_t + e_t$, $cov(r_t, W_t) = 0$, and $e_t \sim N(0, 1^2)$. The OLS averages & bootstrap standard errors are reported.

Table 12
Using log difference approximation for control variables

	N = 100,000 iterations							
	$n = 5$	$n = 20$	$n = 30$	$n = 50$	$n = 100$	$n = 500$	$n = 5000$	$n = 10000$
Actual Return (r_t, from eq. (1))								
$\widehat{\beta}_0$	9.989976	10.00262	9.99809	10.00106	9.998932	9.999751	9.999944	10.00004
$se(\widehat{\beta}_0)$	0.768084	0.244361	0.196144	0.149726	0.104526	0.046571	0.014659	0.010304
$\widehat{\beta}_Z$	4.986003	5.002199	4.998133	4.996505	5.000601	5.000099	5.000095	5.000277
$se(\widehat{\beta}_Z)$	2.21071	0.507734	0.392075	0.295449	0.204727	0.09021	0.028162	0.02018
$\widehat{\beta}_r$	2.016686	1.995824	2.001982	2.002657	2.00006	1.999944	1.999969	1.999724
$se(\widehat{\beta}_r)$	1.949114	0.44796	0.346535	0.259614	0.180332	0.079441	0.024539	0.017629
Log Difference (x_t, from eq. (2))								
$\widehat{\beta}_0$	10.33415	10.29849	10.27203	10.2436	10.21482	10.16863	10.11939	10.11078
$se(\widehat{\beta}_0)$	1.117538	0.389363	0.318987	0.249055	0.185798	0.101624	0.046691	0.038509
$\widehat{\beta}_z$	6.592617	6.835137	6.873831	6.912636	6.95171	7.037271	7.138965	7.155154
$se(\widehat{\beta}_z)$	1.514168	0.452727	0.381766	0.31869	0.260467	0.17049	0.080964	0.067066
$\widehat{\beta}_x$	0.954999	0.675305	0.635234	0.59494	0.549162	0.449282	0.322146	0.299647
$se(\widehat{\beta}_x)$	2.152352	0.620181	0.520415	0.422332	0.335222	0.215757	0.111163	0.094058

$P_t \sim \chi^2{}_5 | [0.5 \leq \overline{r_t} \leq 2]$, $var(e_t|r_t) = \sigma^2$. Data Generating Function (DGF) is $Y_t = 10 + 5Z_t + 2r_t + e_t$, $cov(r_t, Z_t) \neq 0$, and $e_t \sim N(0, 1^2)$. The OLS averages & bootstrap standard errors are reported.

Table 13
The use of $ln(X_{i/t} + 1)$ under homoscedasticity, $var(e_t|r_t) = \sigma^2$

	N = 100,000 iterations							
	$n = 10$	$n = 20$	$n = 30$	$n = 50$	$n = 100$	$n = 500$	$n = 5000$	$n = 10000$
Using $ln(x_i)$ on non-zero subsamples								
$\widehat{\beta}_0$	2.014027	2.005336	1.985509	2.004388	2.001321	1.999733	1.999573	1.999791
$se(\widehat{\beta}_0)$	1.381141	0.863783	0.670689	0.509023	0.352665	0.155488	0.048955	0.034524
$\widehat{\beta}_1$	2.990841	2.997951	3.007542	2.996925	3.000384	3.000276	3.000234	3.00009
$se(\widehat{\beta}_1)$	0.727945	0.455604	0.355857	0.26938	0.186903	0.082509	0.025918	0.018351
Using $ln(x_i + 1)$ on the full sample								
$\widehat{\beta}_0$	5.313541	5.152779	5.092592	5.072529	5.048355	5.02425	5.0225	5.021811
$se(\widehat{\beta}_0)$	1.710399	1.1896	0.96674	0.74952	0.528294	0.23718	0.073545	0.05215
$\widehat{\beta}_1$	1.171221	1.263332	1.293489	1.304573	1.319841	1.331805	1.333019	1.333431
$se(\widehat{\beta}_1)$	0.84485	0.580431	0.47069	0.364772	0.25702	0.115489	0.035723	0.025409

Table 14
The use of $ln(X_{i/t} + 1)$ under heteroscedasticity, $var(e_t|r_t) = \sigma^2{}_i$

	N = 100,000 iterations							
	$n = 10$	$n = 20$	$n = 30$	$n = 50$	$n = 100$	$n = 500$	$n = 5000$	$n = 10000$
Using $ln(x_i)$ on non-zero subsamples								
$\widehat{\beta}_0$	2.227563	1.955088	2.047043	2.013607	1.995693	2.008219	2.003623	1.996921
$se(\widehat{\beta}_0)$	12.89686	7.590617	5.739388	4.19596	2.907235	1.244974	0.395687	0.281454
$\widehat{\beta}_1$	2.88071	3.020849	2.959815	2.991643	3.006319	2.993915	2.997399	3.002468
$se(\widehat{\beta}_1)$	8.467413	5.187395	4.023002	3.000865	2.096051	0.908924	0.289462	0.206641
Using $ln(x_i + 1)$ on the full sample								
$\widehat{\beta}_0$	5.257782	5.083262	5.167984	5.0718	5.041273	5.040372	5.025923	5.015942
$se(\widehat{\beta}_0)$	8.634432	5.828199	4.871704	3.694501	2.598436	1.153893	0.36805	0.26003
$\widehat{\beta}_1$	1.206326	1.294149	1.244652	1.304579	1.326167	1.322169	1.330683	1.33709
$se(\widehat{\beta}_1)$	5.48393	3.673091	3.055456	2.325828	1.639428	0.727015	0.232346	0.16467

Data Generating Function (DGF) is $Y_i = 2 + 3ln(X_i) + e_i$ and $e_i \sim N(0, \sigma^2{}_i)$. The OLS averages & bootstrap standard errors are reported.

again until all values of X_i are strictly positive integers. Next, the DGF $Y_i = 2 + 3ln(X_i) + e_i$, $e_i \sim N(0, 1^2)$, is used to generate Y_i samples. Next, 10 % of the data points in X_i are replaced with zeros. Then, run two different regressions, firstly, using only the 90 % non-zero samples (i.e., subsample) and secondly, using the full sample with $ln(X_i + 1)$ to avoid taking the log of zeros. These are done under homoscedasticity and heteroscedasticity conditions. The simulation results are presented in Table 13 and Table 14. Both under homoscedasticity conditions (Table 13) and heteroscedasticity conditions (Table 14), the OLS estimators of the subsamples are unbiased and consistent for the intercept term and slope. Conversely, the OLS estimators are not only biased for the slope coefficient but also biased for the intercept term. Therefore, these results confirm that the common practice of using $ln(X_{i/t} + 1)$ to avoid a log of zero biases the OLS estimates of the true causal effect, even

Table 15
Empirical Application: Log Difference Approximation, log return.

	$\widehat{\beta}_r$	$\widehat{\beta}_x$	$\lvert\widehat{\beta}_r - \widehat{\beta}_x\rvert$	$\widehat{\beta}_r$	$\widehat{\beta}_x$	$\lvert\widehat{\beta}_r - \widehat{\beta}_x\rvert$	$\widehat{\beta}_r$	$\widehat{\beta}_x$	$\lvert\widehat{\beta}_r - \widehat{\beta}_x\rvert$
Panel A	4.843***	5.535***	0.691	4.851***	5.547***	0.695	5.078***	5.147***	0.069
	(0.823)	(0.836)	(1.173)	(0.825)	(0.838)	(1.176)	(1.359)	(1.331)	(1.902)
$\bar{r}_t$	0.001			0.001			0.001		
$\sigma^2_{r_t}$	0.001			0.001			0.001		
Panel B	1.124***	1.909***	0.785**	1.124***	1.913***	0.788**	1.680***	1.775***	0.094
	(0.262)	(0.288)	(0.390)	(0.263)	(0.289)	(0.391)	(0.481)	(0.459)	(0.665)
$\bar{r}_t$	0.001			0.001			0.001		
$\sigma^2_{r_t}$	0.013			0.013			0.013		
Controls	NO	NO		YES	YES		YES	YES	
N	367	367		367	367		367	367	
F-Statistic	43.853***	34.652***		21.908***	17.301***		14.621***	11.518***	
Adj. R^2	0.107	0.087		0.107	0.087		0.108	0.087	

Note: $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

in large samples. A robust solution to avoid taking the log of zeros is to remove the units, entities, or individuals with zero values from the dataset and use the subsample for estimation. This is because; if the sample is (as if) random, causal effect estimations are guaranteed under the causal effect assumptions. It is important to also mention that the use of $ln(X_{i/t}+1)$ when $X_{i/t}\to\infty$ might not be problematic since $ln(X_{i/t}) - ln(X_{i/t}+1)\to 0$.

3.2. Practical empirical applications

In this section, empirical data analyses are presented for both log difference approximation and $ln(X+1)$ estimations. In addition, the ex-post estimation difference-in-estimates test is performed to show the differences between these common practices and the robust approaches. Table 15 shows the use of log return as a log difference approximation for return under two different situations. In panel A, the effects of using return, $\widehat{\beta}_r$, and log return, $\widehat{\beta}_x$, are presented and their estimates are very close without any significant differences in the estimates, $|\widehat{\beta}_r - \widehat{\beta}_x|$. Guaranteeing that the estimated effects of return are statistically not different from those of log return. This is the outcome when the sample average and variance of returns tend to zero. Particularly, the sample average and variance of the sampled return of Bitcoin are 0.001 and 0.013 respectively. Thus, the results in panel A of Table 15 confirm that when the sample mean and variance of return both tend to zero, the estimated causal effects of returns are not different from that of log return. As such, the log difference approximation of log return can be used in return's stead to correctly estimate the true causal effects of return.

The estimated model is $\Delta V_t = \beta_0 + \beta_i i + \beta_j controls_{jt} + e_t$ where $i\ \varepsilon\ [r_t, x_t]$ and $controls\ \varepsilon\ [covid19, i\times covid19]$. In panel B, the Bitcoin price is transformed using $Price_t = Price_t^{2.8}$ solely to increase the variance of the return while leaving the mean unchanged to investigate the differences in estimate between the log return approximation and the actual return.

In panel B, one of these conditions for log difference approximation fails. That is, while the mean of Bitcoin's return tends to zero, the variance does not. The sample return mean remains 0.001 but the variance is 0.013. Thus, while the mean condition is satisfied, the variance condition is not. The empirical results in panel B confirm that under this situation, the return estimates are statistically different from those of log return. From the simulation analysis so far, it follows that under this scenario, the return estimators are often unbiased and consistent while that of log return is biased, even in large samples.

Similarly, Table 16 shows the estimates from $ln(Patent+1)$, subsample $ln(Patent)$, and subsample $Patent$. Against using $ln(Patent+1)$ and avoid taking the log of zero, the robust

The estimated model is $ESG_t = \beta_0 + \beta_i i + \beta_j controls_{jt} + e_t$ where $i\ \varepsilon\ [Patent_t, lnPatent_t, ln(Patent+1)_t]$. The two differences in estimate test columns are $Q = \left|\widehat{\beta}_p - \widehat{\beta}_{ln(p+1)}\right|$ and $R = \left|\widehat{\beta}_{lnp} - \widehat{\beta}_{ln(p+1)}\right|$.

Approach of taking a none zero subsample for $ln(Patent)$ and $Patent$ show that empirically, these estimates are statistically different. Based on the simulation analysis, the subsample robust approach for $ln(Patent)$ (and $Patent$) produces unbiased and consistent results while that of $ln(Patent+1)$ yields biased results. This is also confirmed by the models' coefficient of determinations which are higher for the robust approaches relative to those of the discussed common practices.

4. Conclusion and recommendations

This article implores researchers to revisit the use of log differencing as an approximation for asset returns, growth rate, and percentage changes for both the interest regressor(s) and control variables. Firstly, log differenced price is never equal to asset returns but can be a good approximation of returns when the returns are sufficiently close to zero i.e., the mean and variance of the returns are very close to zero. Only when these conditions are met will the log difference approximations produce unbiased and consistent causal estimates. However, the log difference estimates are less efficient, less precise, and less accurate. On the other hand, when these conditions are not met, the log difference

Table 16
Empirical Application: The $ln(X+1)$ practice.

	Dependent Variable: ESG									
	$\widehat{\beta}_p$	$\widehat{\beta}_{ln(p)}$	$\widehat{\beta}_{ln(p+1)}$	Q	R	$\widehat{\beta}_p$	$\widehat{\beta}_{ln(p)}$	$\widehat{\beta}_{ln(p+1)}$	Q	R
Estimate	0.241***	2.983***	1.800***	1.559***	1.183***	0.237***	2.368***	1.706***	1.469***	0.662
	(0.018)	(0.348)	(0.245)	(0.246)	(0.426)	(0.020)	(0.421)	(0.290)	(0.291)	(0.511)
Controls	No	No	No			Yes	Yes	Yes		
N	795	795	2932			565	565	1932		
R^2	0.187	0.085	0.018			0.240	0.102	0.028		
Adjusted R^2	0.186	0.084	0.018			0.231	0.091	0.024		
RSE	10.052	10.669	10.229			9.944	10.810	10.246		
F Statistic	182.962***	73.416***	53.965***			25.138***	9.026***	7.861***		

Note: $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

approximations produce biased and inconsistent results which leads to misinformation, invalid inference, incorrect recommendations and policies. Similarly, adding one before taking the natural log of a variable, to avoid taking the natural log of zeros, also produces biased and inconsistent causal estimates. Rather, a robust solution is to use a large enough subsample which correctly estimates the true causal effects under the causal effect assumptions. This study also establishes the benchmark test statistic for both deciding whether (or not) to use the log difference approximation and for adding one to a variable before taking its natural logarithm in a causal regression analysis. However, this benchmark test is an ex-post estimation approach that requires two different estimation results. It will be interesting to develop a robust and sufficient ex-ante estimation test statistic, like Tests I – VI, that sufficiently decides the choice of using log difference approximation at a given significance level.

Generally, logarithmic transformations are handy in empirical data analysis. It could be to make a variable stationary, reduce the variance of a variable, reduce the differences in the magnitude of a variable's data, etc. A key question remains, why do I want to make a variable stationary by log transformation instead of differencing?, why do I want to reduce the differences in the magnitude of a variable?, why do I want to add one to a variable and take a log?, etc. Sometimes, these actions can bias the estimation of the true causal effect, even when the log-transformed variable is a control variable. For example and assuming a variable, patent; adding 1 before taking the log of the patent does not only lie that a firm without patent, $X = 0$, now has a patent, $X^* = 0 + 1$, or that the log of the patent for a firm without a patent remains 0, $ln(X^* = 0 + 1) = X = 0$ but the log of patents for a firm with 2 patents is now approximately 1 patent, $ln(X^* = 2 + 1) = 1.097 < X = 2$, but also biases causal effect estimation. This log-transformed patent has altered the information in the original patent data, $ln(x + 1) - ln(x)$ varies and $ln(x + 1) - ln(x) \to 0$ if $X \to \infty$. A robust solution is to use a large enough subsample, removing all the firms without patents from the sample, for the estimation. Still, the subsample estimates the true causal effect. So, if we understand how the logarithmic transformation alters the information in the data, it becomes easy to decide whether to do it. If the alteration is even or uniform, then, a lesser chance of biasing the causal effect estimates relative to uneven alterations. Of course, $ln(x + 1) \approx ln(x)$ as $X \to \infty$. So, this action has little or no (uneven or unequal) alteration effect on causal estimation when X is large.

The goal is to elicit information from the data and not to alter the information and then elicit the altered information. To the extent the logarithm transformations alter the original variable in a non-uniform manner, it's easier to suspect that the transformation is likely to alter the estimation of the true causal effect. Secondly, if we take the natural logarithm of a variable to reduce the differences between the magnitude of the data points in the variable, we need to remember that while the original magnitude differences vary, the log-transformed magnitude differences remain constant, i.e., $x_2 - x_1 = \alpha$ and $\omega \times x_2 - \omega \times x_1 = \omega \times \alpha$ but $lnx_2 - lnx_1 = ln(\omega \times x_2) - ln(\omega \times x_1) = \beta$. The variability in data is exactly the information we require in data analysis. We want to understand the patterns of these variabilities or variances and their causes. If we restrict these variabilities by our actions, then we have most likely altered the true state of the information embedded in the datasets. A robust solution is to use a subsample, removing outliers with large differences with other data points. The idea of random sampling is that the data is (as if) random and large enough. The idea does not include that a particular data point, firm or individual must be in the sample. Thus, different random samples can consistently produce very similar causal estimates.

Lastly, using logarithmic transformations to make a variable stationarity might be trivial. This is mainly because both stationary and non-stationary variables are capable of estimating the true causal effects, i.e., the right values of the coefficients. However, non-stationary time series commit more type-I errors relative to stationary time series. This is the idea behind using differenced-stationary series to test for unit-root or stationarity as seen in tests like the Dickey-Fuller (DF) test, Augmented Dickey-Fuller (ADF) test, etc. This is because statistical inferences based on a non-stationary series are invalid since their standard errors are biased. Alternatively, the standard errors from a stationary series are unbiased and can make valid inferences under the causal assumptions and parameter distributions. If we realize that we need stationary time series to make valid statistical inferences, why take the log transformation shortcut instead of directly differencing the series and losing one observation? After all, when a time series is differenced, only one observation is lost, which is very trivial if the sample size is large enough. Therefore, only when we answer these questions will the choice be clear.

In conclusion, this study is a gentle reminder and a call for scholars to re-evaluate common practices. Questions that need to be answered before taking these actions might include, not limited to, why do I want to do this?, what effect does this have on the variable?, what effect does it have on estimating the true causal effect?, are there other robust alternatives to achieving the same goal or objective?, etc. These can serve as a roadmap towards correct, valid, powerful, reliable, and influential results that are used for policy formation and implementation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Arnold, C., & Shavelle, R. (1998). Joint confidence sets for the mean and variance of a normal distribution. *American Statistician, 52*, 133–140.

Ausloos, M., Ficcadenti, V., Dhesi, G., & Shakeel, M. (2021). *Benford's laws tests on S&P500 daily closing values and the corresponding daily log-returns both point to huge non-conformity* (p. 574). Physica A: Statistical Mechanics and its Applications.

Bellemare, M., & Wichman, C. (2019). Elasticities and the Inverse Hyperbolic Sine Transformation. *Oxford Bulletin of Economics and Statistics, 28*(1), 50–61.

Bickel, J., & Doksum, K. (1977). *Mathematical Statistics: Basic Ideas andSelected Topics*. San Francisco, CA: Holden-Day.

Blau, B. M., Griffith, T. G., & Whitby, R. J. (2023). Industry regulation and the comovement of stock returns. *Journal of Empirical Finance, 73*, 206–219.

Chen, S., & Gao, J. (2011). Simultaneous specification testing for mean and variance structures in nonlinear time series regression. *Econometric Theory, 27*(4), 792–843.

Chiang, T. C., & Chen, P.-Y. (2023). Inflation risk and stock returns: Evidence from US aggregate and sectoral markets. *The North American Journal of Economics and Finance, 68*.

Choudhari, P., Kundu, D., & Misra, N. (2001). Likelihood ratio test for simultaneous testing ofthe mean and the variance of a normal distribution. *Journal of Statistical Computation andSimulation, 71*, 313–333.

Domenico, F. D., Livan, G., Montagna, G., & Nicrosini, O. (2023). *Modeling and simulation of financial returns under non-Gaussian distributions* (p. 622). Physica A: Statistical Mechanics and its Applications.

Dong, X., & Yu, M. (2023). Does fintech development facilitate firms' innovation? Evidence from China. *International Review of Financial Analysis, 89*.

Duarte-Silva, T., & Kimel, M. T. (2024). Testing excess returns on event days: Log returns vs. dollar returns. *Finance Research Letters, 11*(2), 173–182.

Duran, B., Tsai, W., & Lewis, T. (1976). A class of location-scale tests. *Biometrika, 63*, 173–176.

Fang, V. W., Tian, X., & Tice, S. (2014). Does stock liquidity enhance or impede firm innovation? *The Journal of Finance, 69*(5), 2085–2125.

Halvorsen, R., & Palmquist, R. (1980). The interpretation of dummy variables in semilogarithmic equations. *The American Economic Review, 70*(3), 474–475.

Hao, J., Peng, M., & He, W. (2023). Digital finance development and bank liquidity creation. *International Review of Financial Analysis, 90*.

He, F., Guo, X., & Yue, P. (2024). Media coverage and corporate ESG performance: Evidence from China. *International Review of Financial Analysis, 91*.

Herley, M. D., Orlowski, L. T., & Ritter, M. A. (2023). Asymmetric responses of equity returns to changes in exchange rates at different market volatility levels. *The Journal of Economic Asymmetries, 28*.

Lepage, Y. (1971). A combination of Wilcoxon's and Ansari-Bradley's statistics. *Biometrika, 58*, 213–217.

Lepage, Y. (1973). A table for a combined Wilcoxon Ansari-Bradley statistic. *Biometrika, 60*, 113–116.
Li, J., Nie, H., Ruan, R., & Shen, X. (2024). Subjective perception of economic policy uncertainty and corporate social responsibility: Evidence from China. *International Review of Financial Analysis, 91*.
Liu, C., & Kang, M. (2024). Is the cash-returns relationship risk induced? *The North American Journal of Economics and Finance, 69*(A).
Long, H., Chiah, M., Zaremba, A., & Umar, Z. (2024). Changes in shares outstanding and country stock returns around the world. *Journal of International Financial Markets, Institutions and Money, 90*.
Mullahy, J., & Norton, E. (2023). Why transform Y? The pitfalls of transformed regressions with a mass at zero. *Oxford Bulletin of Economics and Statistics, 86*(2), 417–447.
Neuhäuser, M., Leuchs, A., & Ball, D. (2011). A new location-scale test based on a combination of the ideas of Levene and Lepage. *Biometrical Journal, 53*, 525–534.
Ni, Z., & Wang, L. (2023). The predictability of skewness risk premium on stock returns: Evidence from Chinese market. *International Review of Economics & Finance, 87*, 576–594.
Nick, H. (2023). Linear rescaling to accurately interpret logarithms. *Journal of Econometric Methods, 12*(1), 139–147.
Okorie, D. (2023). Renewable green hydrogen energy: performances amidst global disturbances. *Clean Technologies and Environmental Policy, 26*, 849–873.
Okorie, D., Bouri, E., & Mazur, M. (2024). NFTs versus conventional cryptocurrencies: A comparative analysis of market efficiency around COVID-19 and Russia-Ukraine conflict. *Quarterly Review of Economics and Finance, 95*, 126–151.
Okorie, D., Gnatchiglo, J., & Wesseh, P. (2024). Electricity and cryptocurrency mining: An empirical contribution. *Heliyon, 10*(13).
Okorie, D., & Lin, B. (2023). Cryptocurrency spectrum and 2020 pandemic: Contagion analysis. *International Review of Economics and Finance, 84*, 29–38.
Panagiotidis, T., Papapanagiotou, G., & Stengos, T. (2024). A Bayesian approach for the determinants of bitcoin returns. *International Review of Financial Analysis, 91*.
Park, H. (2015). Simultaneous test for the mean and variance with an application to the statistical process control. *Journal of Statistical Theory and Practice, 9*(4), 868–881.
Pesarin, F., & Salmaso, L. (2010). *Permutation tests for complex data*. New York: Wiley.
Pungaliya, R. S., & Wang, Y. (2023). Machine invasion: Automation in information acquisition and the cross-section of stock returns. *Journal of Financial Markets, 64*.
Rao, C. (1973). *Linear statistical inference and its applications*. New York: John Wiley and Sons.
Rublik, F. (2009). Critical values for testing location-scale hypothesis. *Em Measurement ScienceReview, 9*, 9–15.
Simonato, J.-G., & Denault, M. (2023). Multiperiod portfolio allocation: A study of volatility clustering, non-normalities and predictable returns. *The North American Journal of Economics and Finance, 68*.
Tian, G., Li, B., & Cheng, Y. (2022). Bank competition and corporate financial asset holdings. *International Review of Financial Analysis, 84*.
Tomlinson, M. F., Greenwood, D., & Mucha-Kruczyński, M. (2024). 2 T-POT Hawkes model for left- and right-tail conditional quantile forecasts of financial log returns: Out-of-sample comparison of conditional EVT models. *International Journal of Forecasting, 40*(1), 324–347.
Wong, P. (2023). Explaining intraday crude oil returns with higher order risk-neutral moments. *Journal of Commodity Markets, 31*.
Zhang, D. (2022). Green financial system regulation shock and greenwashing behaviors: Evidence from Chinese firms. *Energy Economics, 111*(c).
Zhang, M., Su, J., Sun, Y., Zhang, W., & Shen, N. (2015). Political connections and corporate diversification: An exploration of Chinese firms. *Emerging Markets Finance and Trade, 51*(1), 234–246.