EJDER
European Journal of Digital Economy Research
e-ISSN : 2717-9230

Volume : 6
Issue : 2
Year : 2025
Pages : 47-57

LEVERAGING REVERSE REGRESSIONS FOR BIAS DIAGNOSIS IN THE DIGITAL ECONOMY DATASETS

Richard MULENGA [a]

[a] ZCAS University, Department of Economics, Lusaka, Zambia, richard.mulenga@zcasu.edu.zm, https://orcid.org/0009-0003-0065-7432.

ABSTRACT:

This paper evaluates reverse regression in simulations and applications motivated by the digital economy data. Data from digital platforms ranging from e-commerce transactions to user-generated content offers vast potential for economic analysis, yet it frequently suffers from measurement errors and endogeneity problems. With digital platforms producing vast amounts of data that are frequently user-created, collected, or compiled, researchers encounter growing difficulties in validating data reliability. The reverse regression provides a unique diagnostic tool set for identifying and correcting biases when the standard assumptions of Ordinary Least Squares (OLS) are not satisfied. This is particularly true in contexts like gig work income reports, online advertising, and consumer trends inferred from internet activities. Based on the simulated digital data of a medium enterprise business digital sales data associated with advertising expenditure reported via Google or Meta dashboards, this study finds that the forward regressions are biased or attenuated. The study therefore recommends that reverse regression involving the digital platform data be applied as a diagnostic and corrective tool set in early-stage econometric diagnostics, especially when robust instrumental variables are unavailable.

Keywords: Reverse regressions, Digital Platform, Digital Economy, Measurement Error, Meta Dashboards.

RECEIVED: 17 June 2025 **ACCEPTED:** 20 November 2025

DOI:
https://doi.org/10.5281/zenodo.18056973

CITE
Mulenga, R., (2025). Leveraging Reverse Regressions for Bias Diagnosis in the Digital Economy Datasets. *European Journal of Digital Economy Research, 6*(2), 47-57. https://doi.org/10.5281/zenodo.18056973

Research Paper

1. INTRODUCTION

The digital economy has changed how data is produced, gathered, and analyzed. Data from digital platforms ranging from e-commerce transactions to user-generated content offers vast potential for economic analysis, yet it frequently suffers from measurement errors and endogeneity problems. The proliferation of digital platforms like Uber and Airbnb has transformed markets, consumer habits, and economic frameworks. These platforms produce extensive data that researchers and policymakers utilize to comprehend market trends, consumer choices, and employment results (Goldfarb & Tucker, 2019). However, data from platforms frequently experience measurement inaccuracies resulting from self-reported information, algorithmic changes, and insufficient validation. Moreover, endogeneity problems often occur from simultaneity and missing variables, undermining causal inference. This study evaluates the capacity of reverse regression as a diagnostic tool to address these econometric issues in the context of the proliferation of digital economy data sets[1].

Reverse regression involves regressing the independent variable of interest against the dependent variable, assessing whether the assumed direction of causality is maintained upon examination (Hansen, 2015; Phillips & Shi, 2018). Initially created for detecting measurement errors (See, for example, Hausman, 2001; Cochrane, 2001) reverse regression has seen restricted use in studies related to the digital economy. This research seeks to fill that gap by investigating the application of reverse regression for identifying measurement error and endogeneity within platform-based datasets. We use actual data from Uber and Airbnb to demonstrate the practical uses and consequences of this diagnostic method.

1.1. Digital Platform-Based Data and Econometric Challenges

Digital platform-based datasets, such as those derived from Uber, Airbnb, facebook, Amazon, Spotify and Google, provide innovative insights but are not originally collected for scholarly research, which presents distinct challenges (Einav & Levin, 2014). Digital platforms enable users to connect, communicate and engage with each other. Additionally, digital platforms provide infrastructure for buying, selling and exchanging goods and services (UNCTAD, 2019). Measurement inaccuracies can arise from platform design elements like rating systems, user behaviors such as self-selection, or data processing techniques including search ranking algorithms. Additionally, endogeneity issues are widespread in digital platform-based data due to the nature of two-sided markets and the presence of feedback loops among variables (Bajari et al., 2015). Measurement errors are common in digital platform data, that often contain variables subject to noise. For instance, ratings can be affected by subjective biases and social desirability effects (Luca, 2016; Nosko & Tadelis, 2015). Location information may lack precision due to reliance on Internet Protocols (IP) based geolocation techniques (Chen et al., 2016). Furthermore, prices are frequently subject to dynamic and algorithmic adjustments, creating ambiguity regarding the actual observed values (Einav & Levin, 2014). Such inaccuracies diminish the reliability of key regressors, availability of robust instrumental variables and can result in biased coefficient estimates in empirical forward (direct) based regression analyses (Bound, Brown, & Mathiowetz, 2001).

In digital marketplaces, simultaneity or endogeneity problems are frequently observed, whereby prices and demand levels influence each other in real time (Athey & Imbens, 2017). Additionally, platform visibility factors, such as search rankings, are impacted and influenced by click-through rates (Ghose & Yang, 2009). Furthermore, unobserved variables such as changes in platform policies or consumer expectations can simultaneously affect both the explanatory variables and the outcomes under consideration (Bajari et al., 2015).

[1] There are various definitions of digital economy offered in the literature. This study adopts the Organization for Economic Cooperation and Development, OECD (2020) and the United Nations Conference on Trade and Development, UNCTAD (2019) definitions. Digital Economy encompasses "all economic activity that is supported by information and communications technologies (ICT), including e-commerce and digitally delivered services" (OECD, 2020). Digital economy entails "the global network of economic and social activities that are enabled by ICT, including e-commerce, digital platforms, and data-driven services" (UNCTAD, 2019).

Although recent developments in causal inference methods, including instrumental variables (IV) and difference-in-differences (DiD), have improved analytical approaches, many studies still face difficulties in establishing credible identification within digital data platforms. The review of extant literature seems to indicate that reverse regression remains an underutilised tool as a diagnostic and corrective tool in the early stages of digital data analyses, and as a corrective tool for bias in financial and economic studies. Additionally, studies employing reverse regressions focusing on the digital economy are scarce. Therefore, this study fills this gap by examining the application of reverse regression as a diagnostic methodology for identifying measurement errors and endogeneity within econometric models, with particular emphasis on the context of the digital economy empirical data analyses (Athey & Imbens, 2017; Goldfarb & Tucker, 2019).

2. LITERATURE REVIEW

2.1. Empirical Literature Review

The concept of using reverse regression dates back to Cochrane (2001) and Hausman (2001), who asserted that assessing how future variables relate to current ones can help in understanding equilibrium relationships and the sources of predictability. Reverse regression has attracted scholarly interest in econometrics due to its effectiveness in identifying classical measurement error, which occurs when an explanatory variable is observed with noise (Hausman, 2001). In traditional contexts, the bias caused by attenuation in ordinary least squares (OLS) estimates can be detected by comparing the coefficients obtained from direct and reverse regressions. A notable discrepancy between these coefficients indicates the likelihood of measurement error (Griliches, 1986). Additionally, reverse regression has been employed to identify simultaneity bias in models where the causal relationship is uncertain (Angrist & Pischke, 2009). For instance, in supply and demand frameworks, reverse regression aids in distinguishing whether price influences quantity or if the causality runs in the opposite direction.

Dufour and Kang (2022) re-evaluate the concept of reverse regression (RR) within the framework of the classical linear regression (CLR) model, emphasizing distributional symmetry and its consequences for hypothesis testing. Although reverse regression has historically been regarded as a special case, often associated with measurement error correction strategies. This study aimed to formalize the RR theoretical foundation and demonstrate its practical and inferential significance. The authors establish a rigorous statistical framework that characterises reverse regression as a mirror counterpart to forward regression, assuming a jointly Gaussian multivariate distribution. Their work advances the literature on reverse regression by providing new theoretical insights. Through a thorough mathematical analysis, Dufour and Kang (2022) reveal profound distributional symmetries connecting reverse and forward regression under standard linear assumptions. Overall, the research affirms reverse regression's role as a valuable addition to the statistician's early-stage econometric analysis toolkit.

Building on this idea, Phillips and Shi (2018) contribute to a growing field of research in econometrics and financial economics focused on identifying, understanding, and dating speculative financial bubbles. They introduce reverse regression as a new method to detect and date stamp explosive financial bubbles. Instead of predicting future prices based on current data, they suggest estimating the explosive behavior by regressing current prices backward onto future prices. Authors assert that RR estimates seem to be less impacted by measurement errors.

Wei and Wright (2013) add to the research on long-term forecasting, especially when it comes to predicting asset returns and macroeconomic indicators using reverse regression methods. These long-horizon regressions, which involve forecasting a variable like stock returns or GDP over several future periods, often face challenges such as small-sample bias, overlapping data points, and poor performance when tested out-of-sample. Wei and Wright investigate whether flipping the regression predicting future values based on current variables can lead to better statistical properties or new insights.

Cready, Hurtt, and Seida (2000) examine a persistent empirical issue in the fields of accounting and finance: the estimation of the relationship between stock returns and accounting earnings. Conventional approaches typically involve forward regressions, where stock returns are regressed on reported earnings or earnings changes to assess their influence on market valuation. Nonetheless, these regressions are susceptible to measurement error, particularly

when the explanatory variable, earnings in this context, is subject to noise or imperfect measurement. The authors investigate whether employing reverse regression can mitigate these measurement errors and produce more accurate inferences regarding the pricing of earnings information. They argue that, according to econometric theory (Hausman, 1978; Griliches & Ringstad, 1970), reverse regression can, under specific assumptions, provide unbiased estimates of the relationship between the variables, especially when the dependent variable (earnings) is measured with less error than the independent variable (returns). Overall, the study asserts that reverse regressions serve as a diagnostic and corrective tool for addressing measurement errors in empirical early-stage diagnostic test analyses.

Goldberger's (1984) publication serves as a fundamental critique and clarification regarding the use of reverse regression within econometric analysis. Goldberger (1984) provides a cautionary perspective on the inappropriate use of reverse regression. Although it does not entirely prohibit its application, the paper advocates for its use in suitable contexts such as diagnostic procedures, exploratory data analysis, or assessments of symmetry, rather than in straightforward structural estimation. The author cautions that any rigorous application of reverse regression in contemporary research must consider the insights and limitations highlighted by this influential work.

Greene (1984) investigated reverse regression methodologies within the framework of wage discrimination analysis, with a focus on labor economics. Greene's (1984) study offers an algebraic and statistical analysis of the consequences of interchanging the roles of dependent and independent variables in discrimination research. He demonstrates that reverse regressions are not merely symmetrical counterparts to forward regressions; the estimated coefficients vary due to differences in variance, covariance structures, and group sizes. The paper provides explicit algebraic derivations illustrating that reverse regression yields estimates of discrimination that are both statistically and economically distinct from those obtained through forward regression, rather than being simple re-expressions of the same underlying relationship. The author explicitly warns against interpreting reverse regression as a dependable approach for measuring discrimination, particularly in the absence of strong assumptions regarding homoscedasticity, normality, and linearity.

Racine and Rilstone (1995) reevaluate the issue of reverse regression in light of criticisms raised by Goldberger (1984) and Greene (1984). The phenomenon known as the "reverse regression paradox" describes the often unexpected and asymmetric outcomes that emerge when the roles of dependent and independent variables are interchanged in a regression analysis. This paradox has generated confusion, particularly within disciplines such as labor economics and finance, where reverse regression techniques have been employed both for diagnostic purposes and inferential analysis.

Schaefer and Visser (2003) contribute to the ongoing scholarly discourse on reverse regression within the framework of employment discrimination analysis, with particular emphasis on wage and salary differentials. Their research serves as a practical addition to the fields of applied econometrics and forensic economics, examining how various regression techniques—namely forward regression, reverse regression, and orthogonal regression produce divergent conclusions regarding the presence and extent of discrimination. The authors contend that orthogonal regression provides a balanced approach by mitigating the arbitrary asymmetry inherent in forward and reverse regression methods.

Cready et al. (2000) examine the application of reverse regression (RR) within the context of financial accounting, with a specific focus on earnings–returns analyses. While conventional approaches typically model stock returns as a dependent variable influenced by earnings through forward regression, RR offers an alternative by treating earnings as a function of returns, citing advantages such as improved interpretability and mitigation of attenuation bias. The researchers utilize both simulated datasets and real financial data to assess and compare the statistical characteristics and interpretative implications of forward versus reverse regression models. Additionally, they investigate how the estimated RR coefficients are affected by various sample selection criteria, definitions of earnings, and different time horizons, aiming to determine whether RR provides more consistent or comprehensible insights. The results reveal that reverse regression yields substantially different coefficient estimates compared to forward

regression, especially in cases where earnings are noisy or contain transient components.

Dereziński and Warmuth (2018) introduce reverse iterative volume sampling (RIVS), an innovative approach for selecting subsets in linear regression analysis. This method intersects the fields of linear regression, randomized algorithms, and subset selection techniques commonly employed in large-scale data applications and kernel methods. It builds upon traditional volume sampling techniques by implementing a reverse, or backward, strategy that enhances both computational efficiency and statistical robustness. Their work extends existing literature by developing a reverse iterative volume sampling—that circumvents the need to compute large determinants, thereby increasing scalability. Dereziński and Warmuth (2018) present a statistically rigorous and computationally efficient methodology for subset selection in linear regression, transforming volume sampling into a reverse elimination process that preserves its desirable statistical qualities while facilitating application to large datasets.

Zeng et al. (2008) examine the issue of univariate calibration in the context of heteroscedasticity. The authors revisit the concept of reversed regression (RR), where concentration is regressed on signal rather than the traditional approach of regressing signal on concentration, and assess whether RR offers advantages under heteroscedastic conditions, particularly in the calibration of instruments. Their research extends prior statistical and chemometric studies by concentrating on heteroscedastic data, a common yet frequently overlooked characteristic in practical calibration scenarios. The findings indicate that RR produces less biased estimates of concentration, especially when the signal is affected by substantial and non-uniform measurement errors, characteristic of heteroscedasticity. Their case study supports the broader conclusion that reverse regression can be statistically advantageous when the objective is to infer an unobserved input variable from noisy output data.

Otero and Baum (2018) examine the robustness and efficacy of unit root tests, specifically the Dickey–Fuller (DF) and augmented Dickey–Fuller (ADF) procedures, by proposing a framework that incorporates both forward and reverse regression approaches. Their goal is to improve the statistical power and dependability of traditional unit root testing methods by analyzing how the directionality of the regression influences inference, particularly in small sample contexts or processes that are close to a unit root. Their results suggest that forward and reverse ADF tests exhibit comparable power under the null hypothesis, yet their performance may differ under certain alternative hypotheses, such as processes nearing a unit root. Researchers and practitioners can leverage these combined forward-reverse ADF testing strategies to enhance diagnostic accuracy of unit root conclusions, especially when working with limited data or data susceptible to measurement errors.

Ash (2014) critically analyzes the application and interpretive challenges associated with reverse regression techniques within the legal fraternity, with a particular focus on employment discrimination cases. The chapter scrutinizes the growing reliance of defendants on reverse regression (RR) as an evidentiary strategy to undermine discrimination claims, highlighting the tendency for such methods to depend on flawed or inappropriate logical assumptions. Ash (2014) emphasizes important methodological concerns, notably that reverse regressions may violate fundamental linear regression assumptions such as exogeneity and proper model specification. The study warns that courts unfamiliar with the statistical limitations of RR may misinterpret its results, leading to potential misjudgments in legal proceedings.

Wei and Wright (2009) examine the methodology for constructing confidence intervals within long-horizon predictive regressions, with particular emphasis on applications in financial economics where the predictability of macroeconomic indicators is evaluated using variables such as dividend-price ratios, interest rates, or inflation rates. The study is driven by the recognized challenge that conventional confidence intervals for long-term coefficients frequently prove to be misleading primarily due to issues such as persistent regressors, small sample biases, and autocorrelation in error terms. To address these concerns, the authors introduce the RR approach. Authors assert that the sampling distribution of the RR-based estimator is less affected by the persistence of predictors and that RR facilitates more precise finite-sample inference compared to traditional forward regression techniques.

Cartwright and Riabko (2024) examine the impact of temporal aggregation, specifically, the transition

from high-frequency data such as daily observations to lower-frequency levels like weekly or monthly, on the accuracy of parameter estimation and the validity of inference within reverse regression models applied to commodity markets. This chapter analyses how such aggregation influences estimation outcomes, revealing that while it may enhance R^2 values, it also tends to inflate standard errors and diminish the reliability of inferential conclusions, ultimately resulting in less accurate forecasts. The research highlights the importance of exercising caution when employing aggregated data for reverse regression analyses in financial contexts.

Chen (2011) investigates the performance of forward (direct) and reverse regressions in estimation, employing both ordinary least squares (OLS) and instrumental variables (IV) methodologies. The research focuses on analyzing returns to scale and technological progress within the U.S. manufacturing sector over a span of approximately fifty years. Operating within an error-in-variables (EIV) framework where both input and output growth rates are subject to noise measurement, the study demonstrates that OLS estimates, whether direct or reverse, are inconsistent, with the reverse OLS generally exhibiting greater precision under the assumption of normality. Overall, the study underscores the significant bias present in OLS forward and reverse regressions due to measurement error and confirms that reverse IV estimation effectively corrects for this bias, ensuring consistency across both regression orientations.

2.2. Theoretical Underpinnings of Reverse Regressions

In describing the theoretical underpinnings, the study adapts the notations in Cochrane (2001) Hausman (2001) and Hanssen (2015).

Suppose the variables $(\boldsymbol{y}, x)$ are jointly normally distributed. Consider the best predictor of y given $\boldsymbol{x}$

$$y = \boldsymbol{x}'\boldsymbol{\beta} + e \quad (1)$$
$$\beta = (\mathbb{E}(\boldsymbol{x}\boldsymbol{x}'))^{-1}\,\mathbb{E}(\boldsymbol{x}y)$$

Given that the error e is a linear transformation of the normal vector$(y, \boldsymbol{x})$, it follows that $((e, \boldsymbol{x})$ is jointly normal.

Additionally, given that $\mathbb{E}(\boldsymbol{x}e) = 0$ are jointly normal and uncorrelated, this means that they are also independent. Independence in this context implies that:

$\mathbb{E}(e|\boldsymbol{x}) = \mathbb{E}(e) = 0$ and

$$\mathbb{E}(e^2|x) = \mathbb{E}(e^2) = \sigma^2 \quad (2)$$

Equation 2 denotes properties of a homoscedastic linear conditional expectation function, CEF (Hansen, 2015). Given that $(y, \boldsymbol{x})$ are jointly normally distributed, they satisfy a normal linear CEF.

$$y = \boldsymbol{x}'\boldsymbol{\beta} + e \quad (3)$$

Where $e \sim N(0, \sigma^2)$ is independent of $\boldsymbol{x}$

The theoretical discussions in equations 1 and 2 represent the 'traditional' motivation for the linear CEF models. However, it is contended that this motivation has limited merit in econometric applications given that, on the whole, economic and finance data is usually non-normally distributed (Hausman, 2001; Hansen, 2015).

Consider the classical linear model (CLM):

$$Y_i = \beta \boldsymbol{X}_i + \varepsilon_i \quad (4)$$

Suppose $\boldsymbol{X}$ is measured with error:

$$\boldsymbol{X}_i = \boldsymbol{X}_i + v_i \quad (5)$$

Where v_i is the classical measurement error or attenuation bias. This implies that the ordinary least squares (OLS) estimator of β is biased towards zero.[2] In high-dimensional contexts, regularized variants of ordinary least squares, such as ridge regression, penalize the coefficients to draw them closer to zero, thereby decreasing variance. Although this process does not

[2] Measurement error in explanatory variables is a well-recognized source of attenuation bias, causing the ordinary least squares (OLS) estimator of β to be biased towards zero (Greene, 2010; Hansen, 2015)

Suppose the true model is: $y = \beta x^* + \varepsilon$, but instead of observing x^*, we observe $x = x^* + u$, where u is classical measurement error. Estimating $y = \beta x + \varepsilon'$ using OLS leads to bias:

$\widehat{\beta_{OLS}} \rightarrow \lambda\beta$ where; $0 < \lambda < 1$. This causes systematic underestimation of the effect of x^* on y (Wooldridge, 2010; Hansen, 2015). Furthermore, regularization methods such as ridge regression inherently introduce a shrinkage bias by design, which pulls parameter estimates closer to zero in order to reduce variance. The estimated β values are systematically pulled toward zero to minimize mean squared error (MSE).

$\widehat{\beta_{ridge}} = (X'X + \lambda I)^{-1}X'y$ as $\lambda > 0$, this introduces shrinkage and a bias toward zero (Hastie, Tibshirani, & Friedman, 2009). In environments characterized by high levels of noise or when omitted variables are correlated with the included regressors, similar patterns of bias may also arise (Einav & Levin, 2014).

constitute 'bias' in the traditional sense, the estimated beta coefficients are deliberately biased toward zero to optimize the mean squared error, MSE (Hausman, 2001).

Given these sources of attenuation bias, reverse regression (RR) mitigates the attenuation bias by flipping the dependent and independent variables:

$$\boldsymbol{X}_i = \gamma_1 \boldsymbol{Y}_i + \eta_i \qquad (6)$$

If X is measured with error, the reverse regression slope γ_1 can be compared to the inverse of β (Cochrane ,2001). The slope γ_1 will differ in magnitude and direction depending on the presence and nature of the measurement error or endogeneity.

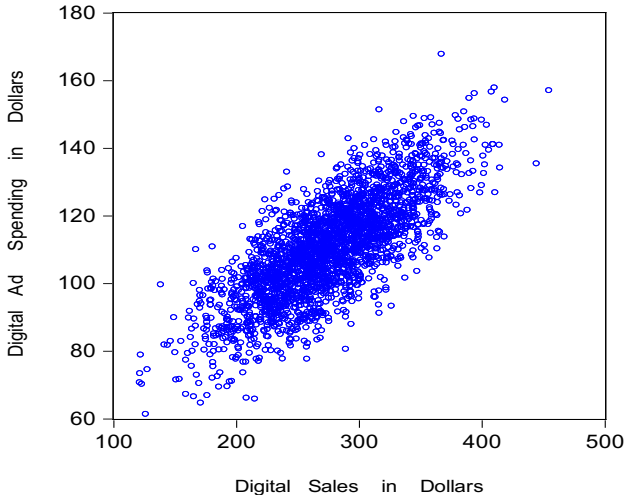
3. METHODOLOGY

3.1. Data Descriptions

The dataset employed in this research was synthetically created to mimic the typical features of digital economy data influenced by measurement errors. Specifically, the data represents a hypothetical online company focused on digital sales. The study envisions a scenario of a medium-sized enterprise business digital sales data associated with advertising expenditure reported via Google or Meta dashboards. The study evaluates the impact of digital advertising expenditure on digital sales. This dataset includes the t*rue digital advertising spending unobserved in practice* (following the model in Zeng et al.2008), the r*eported digital advertising spending with measurement error*, the *digital sales*, measurement error (*added to* the true and reported digital advertising spending) and t*he random error (added to the digital sales*. Utilizing conventional econometric simulation techniques[3], the true independent variable; reported digital advertising expenditure, was sampled from a normal distribution. Subsequently, measurement errors were systematically incorporated to replicate the noisy reporting typical of digital datasets. The dependent variable, digital sales, was generated as a linear function of the true independent variable with the addition of random error. This methodology establishes a controlled setting to assess the efficacy of reverse regression as a diagnostic and corrective instrument in the presence of realistic data noises frequently encountered in digital economy data analyses (Wei & Wright,2013; Ryan, & Yang, 2015).

Figure 1 shows the scatter plot for digital sales and digital (reported advert) advertising expenditure. The plot clearly shows there is a positive correlation between online reported advertising and digital sales.

Figure 1. Scatter Plot Digital Advertising vs Digital Sales



Source: Author's elaboration on data

The relatively positive correlation between online advertising expenditure and digital sales can be explained from the perspective that online adverts help make products more visible, raise consumer awareness, and attract more visitors to online stores where purchases happen. Several important factors explain this positive correlation. First, online adverts boost awareness of products and strengthen brand recognition, making consumers more likely to consider buying. When people see adverts on social media, search engines, or websites, they learn about product features, benefits, and special offers, which can spark interest and the desire to buy (Grewal et al., 2020). Second, targeted advertising allows companies to reach the right audiences, who are more likely to purchase. Digital platforms analyze user data and browsing habits to show adverts to individuals whose behavior indicates a higher chance of buying. This focused approach reduces wasted advertising expenditure and makes campaigns more effective (Lambrecht & Tucker, 2013). Third, interactive and personalized adverts

[3] The study employed the Scenario Analysis to generate the hypothetical data following the Kydland & Prescott (1982) model. The Scenario Analysis is supplemented by the static microsimulation technique, given that the data generated are firm-level microeconomic data (Li & O'Donoghue, 2013).

help engage customers more deeply, which can lead to more conversions and potential increases in purchases. Features like clickable ads, product demos, and real-time feedback enable users to interact with products before deciding to buy, building trust and increasing the chances of purchase (Chaffey & Ellis-Chadwick, 2019). Lastly, the quick and easy nature of digital platforms enhances the impact of advertising. For instance, after seeing an ad, consumers can immediately click to view a product page, read reviews, and complete their purchase. This smooth process shortens the decision-making time and encourages more sales (Stephen, 2016).

3.2. Simulation Strategy

Model 1: Direct OLS Regression

$$Y = \beta_0 + \beta_1 \boldsymbol{X} + u \quad (7)$$

Model 2: Reverse Regression

$$\boldsymbol{X} = \gamma_0 + \gamma_1 \boldsymbol{Y} + \eta \quad (8)$$

If X is measured with error, the reverse regression slope, γ_1 can be compared to the inverse of β (Cochrane, 2001). That is, the bias indicator is estimated as:

$$Bias = \left|\hat{\beta}_1\right| / \left|\hat{\gamma}_1\right| \quad (9)$$

This ratio serves as a diagnostic measure of the degree and direction of potential bias or measurement error.

A prori expectations:

(i) When $Bias = 0$; forward and reverse regressions agree, however,

(ii) When $Bias > 0$; the forward estimate is attenuated or biased towards zero and differs from the reciprocal of the regression slope.

The measurement error is defined as: $X_i = X_i + v_i^*$, where $v_{i\sim}N(0, \sigma^2)$.

In summary, the simulation strategy involves running both direct and reverse regressions on simulated data, then comparing the estimated slopes and their statistical significance.[4]

4. RESULTS AND DISCUSSIONS

Table 1 reports summary statistics of the simulated data.

Table 1. Summary Statistics

	DIG_SAL	TRUE AD_SPEND	REPORTED_AD_SPEN	RANDOM_ERROR_SALE	MEASUREMENT _ERROR
Mean	275.1278	110.5114	110.2989	-1.150864	-0.212552
Median	274.2002	110.413	110.7415	-1.382851	-0.067832
Max.	454.4609	167.791	166.8657	105.8717	39.26238
Min.	121.3624	61.38099	54.00618	-110.651	-31.76704
Std. Dev.	48.21366	14.81024	17.81859	30.87028	10.05153
Sum	704327	282909.3	282365.2	-2946.212	-544.1321
Sum Sq. Dev.	5948540	561299.6	812487.8	2438661	258543.9
Obs.	**2560**	**2560**	**2560**	**2560**	**2560**

Notes: Dig_Sal is Digital Sales, Advert_Spend is digital Advertising Spending or expenditure, Reported_Ad_Spend is reported online Advertising Spending, Random_Error_Sale is Random error in Sales.

Table 2 reports the correlation relationships among the variables in the sample. From Table 2, it seems both the true advertising spending (True_Ad_spending) and the reported spending (Reported_AD_Spending) indicates a relatively strong positive correlation between digital expenditure and digital sales. This observation is in tandem with the results obtained in the scatter plot in Figure 1, where it is found that there is a positive correlation between online advertising expenditure and digital sales. The relatively strong positive correlation shown in Table 2 can be explained by the fact that online adverts generally help make products more visible, raise consumer awareness, and attract more visitors online or digital stores, which increases the probability of making purchases (Chaffey & Ellis-Chadwick, 2019; Grewal et al., 2020).

[4] If suitable instrumental variables (IV) existed, the diagnostics would be compared with IV results obtained via instrumental variable regressions taking this format: $Y = \delta_0 + \delta_1 \hat{X} + \varepsilon$, where $\hat{X}$ is predicted from instrument Z.

Table 2. Correlation Matrix

	DIG_SAL	AD_SPEND	REPORTED_AD_SPE	RANDOM_ERROR_SA	MEASUREMENT_ERR
DIG_SAL_	1				
TRUE_AD_SPEND	0.768	1			
REPORTED_AD_SPEND	0.636	0.826	1		
RANDOM_ERROR_	0.640	0.004	0.003	1	0.005
MEASUREMENT_ERRO	-0.004	-0.009	0.556	0.005	1

Table 3 reports a summary of results obtained in the forward (direct) and reverse (backward) regressions employing equations 7 and 8. The variables were in first difference format following the non-stationarity test results of each variable in the unit root test diagnostics (not reported here).

Table 3. Summary Results of Forward and Reverse Regressions

	Coeff	Std.Error	t-Stat	P_value	Mean Dependent Var
Model 1: Forward Regressions (Method-Least Squares)					
Model 1A: Dependent variable- Sales with random error					
Independent Variable- True-Ad-Spending	4.972**	0.204	24.387	0.00	273.976
Model 1B: Dependent variable- Sales with random error					
Independent Variable- Reported_Ad_Spending	3.174**	0.146	-16.902	0.00	273.976
Model 2: Reverse Regressions					
Model 2A: Independent variable- Sales with random error					
Dependent Variable-True_Ad_Spending	2.706**	0.004	23.903	0.00	110.086
Model 2B:					
Dependent Variable- Reported_Ad_Spending	1.606**	0.007	16.262	0.00	110.086

Source: Author's elaboration on simulated data. ** denotes statistical significance at 5%.

Results in Table 3 indicate that both true and reported digital advertising spending have positive, statistically significant effects on digital sales in the forward regressions (model 1). Specifically, ceteris paribus, an increase of 1 unit in true digital/online advertising expenditure causes an increase in digital sales of 4.97 %, and a unit increase in reported online advertising expenditure increases digital sales by 2.47%.

Similarly, in model 2, the reverse regressions indicate that both the true and reported digital expenditures have a positive and significant effect on digital sales. Specifically, holding other factors constant, the true and reported digital expenditures increase digital sales by 2.71% and 3.12%, respectively.

The positive significant effect of both the true and reported digital expenditures on digital sales can be explained in part by the fact that online adverts help make products more visible, raise consumer awareness, and attract more visitors to online stores where purchases happen. Online adverts boost awareness of products and strengthen brand recognition, making consumers more likely to consider buying. When people see adverts on social media, search engines, or websites, they learn about product features, benefits, and special offers, which can spark interest and the desire to buy (Grewal et al., 2020). Features like clickable ads, product demos, and real-time feedback enable users to interact with products before deciding to buy, building trust and increasing the chances of purchase (Chaffey & Ellis-Chadwick, 2019).

4.1. Determining the Attenuation Bias (estimating the bias indicator)

This section employs equation 9 to estimate the degree and direction of bias (or measurement error. That is;

$$Bias = |\hat{\beta}_1| / |\hat{\gamma}_1|$$

Where $\hat{\beta}_1$ is the slope coefficient for the forward regressions,$\hat{\gamma}_1$ is the slope coefficient for the reverse regressions. And the a-priori expectations are:

When $Bias = 0$; forward and reverse regressions agree, however,

When $Bias > 0$; the forward estimate is attenuated or biased towards zero and differs from the reciprocal of the regression slope.

Model 1A compared with Model 2A:

$$Bias = |\hat{\beta}_1| / |\hat{\gamma}_1| = \frac{4.97}{2.71} = 1.83$$

Model 1B compared with Model 2B:

$$Bias = |\hat{\beta}_1| / |\hat{\gamma}_1| = \frac{3.17}{1.61} = 1.97$$

The bias ratio is positive in both model 1 and model 2. This implies that the forward (direct) estimates obtained through forward regressions are biased towards zero (attenuated).

CONCLUSION

This study presents reverse regression as a diagnostic tool to address these econometric issues in the context of the digital economy. The study is conducted in the backdrop of the fact that digital platforms produce extensive data that researchers and policymakers utilize to comprehend market trends, consumer choices, and employment. However, data from these platforms frequently experience measurement errors resulting from self-reported information, algorithmic changes, and insufficient validation. Moreover, endogeneity problems often occur from simultaneity and missing variables, undermining causal inference. This study presents reverse regression as a diagnostic tool to address these econometric issues in the context of digital economy. The dataset employed in this study was synthetically created to mimic the typical features of digital economic data influenced by measurement errors. Specifically, the data represents a hypothetical online company focused on digital sales. The study evaluates the impact of digital advertising expenditure on digital sales. Findings indicate a relatively strong positive correlation between digital advertising expenditure and digital sales. Additionally, both forward and reverse regressions indicate a significant positive effect of online advertising spending on digital sales. Specifically, in the forward regressions (model 1), ceteris paribus, an increase of 1 unit in true digital/online advertising expenditure causes an increase in digital sales of 4.97 %, and a unit increase in reported online advertising expenditure increases digital sales by 2.47%. In the reverse regressions (model 2), ceteris paribus, the true and reported digital expenditures increase digital sales by 2.71% and 3.12%, respectively. The bias indicator shows that the bias in both models is positive. This shows that the forward regressions are biased or attenuated. The findings of this study corroborate the results obtained by Chen (2011), Weigh and Wright (2009) and Zeng et al (2008). These studies assert that the findings indicate that reverse regressions produce less biased estimates relative to forward(direct) regressions. Given these findings, the study recommends that reverse regression involving the digital economy data be applied as a diagnostic and corrective tool set in early-stage econometric diagnostics, especially when robust instrumental variables are unavailable.

REFERENCES

Angrist, J. D., & Pischke, J. S. (2009). Mostly harmless econometrics: An empiricist's companion. *Princeton University Press*.

Ash, A. S. (2014). The perverse logic of reverse regression. In Statistical methods in discrimination litigation. *Taylor & Francis*.

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2), 3–32.

Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement error in survey data. In *Handbook of Econometrics* (Vol. 5, pp. 3705–3843). Elsevier.

Cartwright, P. A., & Riabko, N. (2024). Temporal aggregation and the estimation of reverse regressions for commodities market models. *In Handbook of investment analysis, portfolio management, and financial derivatives* (Vol. 4, Chapter 8). World Scientific.

Chaffey, D., & Ellis-Chadwick, F. (2019). Digital marketing (7th ed.). *Pearson Education*.

Chen, X. (2011). Increasing returns to scale in U.S. manufacturing industries: Evidence from direct and reverse regression [Working paper]. *Bureau d'Économie Théorique et Appliquée*, University of Strasbourg. https://core.ac.uk

Chen, Y., Pavlov, D., & Canny, J. (2016). Large-scale behavioral targeting. *In Proceedings of the 15th ACM SIGKDD,* International Conference on Knowledge Discovery and Data Mining (pp. 209–218).

Cochrane, J. H. (2001). Asset pricing. *Princeton University Press.*

Cready, W. M., Hurtt, D. N., & Seida, J. A. (2000). Applying reverse regression techniques in earnings–return analyses. *Journal of Accounting and Economics*, 30(1), 25–45.

Dereziński, M., & Warmuth, M. K. (2018). Reverse iterative volume sampling for linear regression. *Journal of Machine Learning Research*, 19, 1–41.

Dufour, J.-M., & Kang, B. (2022). Reverse regressions, symmetry and test distributions in linear models. *Journal of Quantitative Economics*, 20(1), 1–25

Einav, L., & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy*, 14(1), 1–24.

Ghose, A., & Yang, S. (2009). An empirical analysis of search engine advertising: Sponsored search in electronic markets. *Management Science*, 55(10), 1605–1622.

Goldfarb, A., & Tucker, C. (2019). Digital economics. *Journal of Economic Literature*, 57(1), 3–43.

Greene, W. H. (1984). Reverse regression: The algebra of discrimination. *Journal of Business & Economic Statistics,* 2(1), 1–8.

Greene, W.H (2010). Econometrics Analysis. 6th edition. Springer.

Grewal, D., Hulland, J., Kopalle, P. K., & Karahanna, E. (2020). The future of technology and marketing: A multidisciplinary perspective. *Journal of the Academy of Marketing Science*, 48(1), 1–8.

Hansen, L. P. (2015). Uncertainty outside and inside economic models. *Journal of Political Economy*, 122(5), 945–987.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: *Data mining, inference, and prediction* (2nd ed.). Springer.

Hausman, J. A. (2001). Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *Journal of Economic Perspectives*, 15(4), 57–67.

Kydland, F. E., & Prescott, E. C. (1982). Time to build and aggregate fluctuations. *Econometrica,* 50(6), 1345–1370.

Lambrecht, A., & Tucker, C. (2013). When does retargeting work? Information specifically in online advertising. *Journal of Marketing Research*, 50(5), 561–576.

Li, J., & O'Donoghue, C. (2013). A survey of dynamic microsimulation models: Uses, model structure and methodology. *International Journal of Microsimulation*, 6(2), 3–55.

Luca, M. (2016). Reviews, reputation, and revenue: *The case of Yelp.com*. Harvard Business School NOM Unit Working Paper No. 12-016.

Nosko, C., & Tadelis, S. (2015). The limits of reputation in platform markets: *An empirical analysis and field experiment*. NBER Working Paper No. 20830.

OECD. (2020). Measuring digital transformation: *A roadmap for the future*. OECD Publishing.

Otero, J., & Baum, C. F. (2018). Unit-root tests based on forward and reverse Dickey–Fuller regressions. *The Stata Journal*, 18(1), 145–164.

Phillips, P. C. B., & Shi, S.-P. (2018). Financial bubble implosion and reverse regression. *Econometric Theory*, 34(3), 705–753.

Stephen, A. T. (2016). The role of digital and social media marketing in consumer behavior. *Current Opinion in Psychology*, 10, 17–21.

UNCTAD. (2019). Digital Economy Report 2019: Value creation and capture – Implications for developing countries. *United Nations*.

Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data (2nd ed.). *MIT Press.*

Zeng, Q. C., Zhang, E., & Tellinghuisen, J. (2008). Univariate calibration by reversed regression of heteroscedastic data: A case study. *The Analyst*, 133(10), 1340–1345.