




Research Article

Representation of Omitted Variable Bias with the Total Derivative Method

Tolga Omay^{1*}, Zeynep Elitaş² 

¹Department of Economics, Atilim University, 06830 Ankara, Turkey

²Department of Labor Economics and Industrial Relations, Anadolu University, 26470 Eskişehir, Turkey
Email: omay.tolga@gmail.com

Received: 31 May 2023; **Revised:** 26 June 2023; **Accepted:** 11 August 2023

Abstract: This study aims to provide an understanding of the concept of omitted variable bias through the total derivative method. This novel approach that is often overlooked could bring a new perspective to statisticians, econometricians, or researchers in neighboring disciplines such as social sciences, management, or economics. In order to complement this mathematical method, the study also employs graphical representations. By doing so, we provide a detailed walkthrough of the total derivative method, its visual depiction, and its application to the omitted variable bias. We believe that this approach can enhance the understanding of regression analysis and foster a deeper connection between mathematics and econometrics. Overall, this study can contribute to the development of new theoretical foundations using the total differential method in this context.

Keywords: omitted variable bias, total differential, total derivative, multivariate regression

MSC: 62H12, 34A30, 35A25

1. Introduction

In the initial stage of university education, particularly in economics and other social sciences, difference operators are commonly used to explain various concepts. This phenomenon is evident, for instance, in the introduction to first-year economics courses, where the concept of elasticity is explained using a difference operator. Subsequently, in the second year of study, it is emphasized that these differences converge to zero, resulting in the transition from difference operators to derivative operators. At this stage, the derivative formulas for point elasticity are introduced instead of arc elasticity. It is noteworthy, however, that a comparable approach is not typically observed in undergraduate statistics or econometrics courses.

By highlighting this disparity, our study aims to bridge the gap between the application of difference operators in basic economic concepts and their extension to derivative operators in a statistical context. We seek to demonstrate the relevance and significance of utilizing the total derivative method in statistical analysis, thereby facilitating a deeper understanding of the subject matter. This contribution is particularly valuable as it enhances the integration of mathematical principles and statistical techniques in the field of economics and related disciplines.

The total differential or derivative concept is not an analytical tool in statistics or econometrics textbooks. In

general, statistics or econometrics is taught using algebra at the undergraduate level, while in higher-level courses, it is taught with matrix algebra. In order to understand multivariate regression analysis, many algebraic derivations are utilized at the undergraduate level. In addition, various examples are given for multivariate regression analysis in the undergraduate-level books. The meaning of partial regression parameters is explained with these examples and derivations. The primary purpose of these efforts is to explain that each partial regression parameter contributes to dependent variables while other parameters are constant. Model misspecification describing multivariate regression and omitted variables is analyzed using different analytical techniques at the undergraduate and graduate levels. However, there is a lack of explanation of the multiple regression partial parameters or omitted variable bias through the total differential or derivative in statistics or econometrics materials in general. It seems that econometrics, or mathematical statistics, lacks a critical tool in this respect. This study aims to develop a method for representing omitted variable bias with the total derivative, which fills this gap in mathematical statistics. With this contribution, we believe it is evident that techniques used in the continuous domain can also provide significant insights into the discrete domain. In addition, we provide a graphical representation that enriches the understanding of the relationship between these two theoretical relations with each other. This attempt is expected to shed light on other statistics or econometric issues from a different perspective and lead to new theoretical foundations. It is far beyond the scope of this article to include all the efforts made to understand the information content of the partial regression parameter in multivariate regression analysis. For this reason, it is decided to discuss the subject by being limited to the descriptions of Gujarati and Porter [1] at the undergraduate level. They have explained the omitted variable bias using algebra techniques.

The paper is organized as follows: Section 2.1 gives a detailed algebraic explanation of multivariate regression and omitted variable bias in line with the definitions of Gujarati and Porter [1]. In Section 2.2, the concept of the total differential and derivative is argued in accordance with the Chiang and Wainwright [2] context. In Section 3, the concepts of omitted variable bias and total differential/derivative will be discussed. This section provides an understanding of obtaining the omitted variable bias with the total derivative. Furthermore, the auxiliary graphic developed for a better understanding of the subject is presented. Finally, the concluding section stresses the importance of this new proposed method.

2. Theoretical explanation and background

2.1 Model misspecification: Multivariate regression and omitted variable bias

First, the derivation of the omitted variable bias will be shown with the deviation model from the mean. For this purpose, the multivariate regression model in which two variables are included is considered as follows:

$$Y_i = \beta_1 + \beta_2 X_{i,2} + \beta_3 X_{i,3} + u_i \quad (1)$$

$$Y_i = \beta_1 + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \dots + \beta_{p-1} X_{i,p-1} + \beta_p X_{i,p} + u_i \quad (1')$$

$$Y_i = \beta_1 + \sum_{j=2}^p \beta_j X_{i,j} + u_i \quad (1'')$$

where i denotes the entities or cross-sectional data. Equation (1') is the generalized version, and equation (1'') is the generalized compact version. For the deviation form, all the variables in the equation are subtracted from their means:

$$\begin{aligned} (Y_i - \bar{Y}) &= \beta_1 (X_{i,0} - \bar{X}_0) + \beta_2 (X_{i,2} - \bar{X}_2) + \beta_3 X_{i,3} (X_{i,3} - \bar{X}_3) + (u_i - \bar{u}) \\ (Y_i - \bar{Y}) &= \sum_{j=1}^p \beta_j (X_{i,j-1} - \bar{X}_{j-1}) + (u_i - \bar{u}) \\ (Y_i - \bar{Y}) &= \beta_1 (X_{i,0} - \bar{X}_0) + \beta_2 (X_{i,2} - \bar{X}_2) + \dots + \beta_p X_{i,p} (X_{i,p} - \bar{X}_p) + (u_i - \bar{u}) \end{aligned} \quad (2)$$

Since $y_i = (Y_i - \bar{Y})$, $X_{i,0} - \bar{X}_0 = 0$, $X_0 = (1, 1, 1, \dots, 1)$, $\bar{X}_0 = 1$, $x_{i,2} = (X_{i,2} - \bar{X}_2)$, and $x_{i,3} = (X_{i,3} - \bar{X}_3)$, we obtain the

following equation:

$$\begin{aligned}
 y_i &= \beta_2 x_{i,2} + \beta_3 x_{i,3} + (u_i - \bar{u}) \\
 y_i &= \beta_2 x_{i,2} + \beta_3 x_{i,3} + \dots + \beta_{p-1} x_{i,p-1} + \beta_p x_{i,p} + (u_i - \bar{u}) \\
 y_i &= \sum_{j=2}^p \beta_j x_{i,j} + (u_i - \bar{u})
 \end{aligned} \tag{3}$$

We multiply equation (3) by $x_{i,2}$, and then $x_{i,3}$ to get the normal equations:

$$\begin{aligned}
 y_i x_{i,2} &= \beta_2 \underbrace{x_{i,2} x_{i,2}}_{x_{i,2}^2} + \beta_3 x_{i,3} x_{i,2} + x_{i,2} (u_i - \bar{u}) \\
 y_i x_{i,3} &= \beta_2 x_{i,2} x_{i,3} + \beta_3 \underbrace{x_{i,3} x_{i,3}}_{x_{i,3}^2} + x_{i,3} (u_i - \bar{u}) \\
 y_i x_{i,2} &= \beta_2 \underbrace{x_{i,2} x_{i,2}}_{x_{i,2}^2} + \beta_3 x_{i,3} x_{i,2} + \dots + \beta_{p-1} x_{i,p-1} x_{i,2} + \beta_p x_{i,p} x_{i,2} + x_{i,2} (u_i - \bar{u}) \\
 y_i x_{i,3} &= \beta_2 x_{i,2} x_{i,3} + \beta_3 \underbrace{x_{i,3} x_{i,3}}_{x_{i,3}^2} + \dots + \beta_{p-1} x_{i,p-1} x_{i,3} + \beta_p x_{i,p} x_{i,3} + x_{i,3} (u_i - \bar{u}) \\
 y_i x_{i,p-1} &= \beta_2 x_{i,2} x_{i,p-1} + \beta_3 x_{i,3} x_{i,p-1} + \dots + \beta_{p-1} \underbrace{x_{i,p-1} x_{i,p-1}}_{x_{i,p-1}^2} + \beta_p x_{i,p} x_{i,p-1} + x_{i,p-1} (u_i - \bar{u}) \\
 y_i x_{i,p} &= \beta_2 x_{i,2} x_{i,p} + \beta_3 x_{i,3} x_{i,p} + \dots + \beta_{p-1} x_{i,p-1} x_{i,p} + \beta_p \underbrace{x_{i,p} x_{i,p}}_{x_{i,p}^2} + x_{i,p} (u_i - \bar{u})
 \end{aligned} \tag{4}$$

$$\begin{aligned}
 y_i x_{i,2} &= \sum_{j=2}^p \beta_j x_{i,j} x_{i,2} + x_{i,2} (u_i - \bar{u}) \\
 y_i x_{i,3} &= \sum_{j=2}^p \beta_j x_{i,j} x_{i,3} + x_{i,3} (u_i - \bar{u}) \\
 &\vdots \\
 &\vdots \\
 y_i x_{i,p-1} &= \sum_{j=2}^p \beta_j x_{i,j} x_{i,p-1} + x_{i,p-1} (u_i - \bar{u}) \\
 y_i x_{i,p} &= \sum_{j=2}^p \beta_j x_{i,j} x_{i,p} + x_{i,p} (u_i - \bar{u})
 \end{aligned}$$

Get the sum of both sides of the equations:

$$\begin{aligned}
 \sum y_i x_{i,2} &= \beta_2 \sum x_{i,2}^2 + \beta_3 \sum x_{i,3} x_{i,2} + \sum x_{i,2} (u_i - \bar{u}) \\
 \sum y_i x_{i,3} &= \beta_2 \sum x_{i,2} x_{i,3} + \beta_3 \sum x_{i,3}^2 + \sum x_{i,3} (u_i - \bar{u}) \\
 \sum y_i x_{i,2} &= \beta_2 \sum x_{i,2}^2 + \beta_3 \sum x_{i,3} x_{i,2} + \dots + \beta_{p-1} \sum x_{i,p-1} x_{i,2} + \beta_p \sum x_{i,p} x_{i,2} + \sum x_{i,2} (u_i - \bar{u}) \\
 \sum y_i x_{i,3} &= \beta_2 \sum x_{i,2} x_{i,3} + \beta_3 \sum x_{i,3}^2 + \dots + \beta_{p-1} \sum x_{i,p-1} x_{i,3} + \beta_p \sum x_{i,p} x_{i,3} + \sum x_{i,3} (u_i - \bar{u}) \\
 \sum y_i x_{i,p-1} &= \beta_2 \sum x_{i,2} x_{i,p-1} + \beta_3 \sum x_{i,3} x_{i,p-1} + \dots + \beta_{p-1} \sum x_{i,p-1}^2 + \beta_p \sum x_{i,p} x_{i,p-1} + \sum x_{i,p-1} (u_i - \bar{u}) \\
 \sum y_i x_{i,p} &= \beta_2 \sum x_{i,2} x_{i,p} + \beta_3 \sum x_{i,3} x_{i,p} + \dots + \beta_{p-1} \sum x_{i,p-1} x_{i,p} + \beta_p \sum x_{i,p}^2 + \sum x_{i,p} (u_i - \bar{u})
 \end{aligned} \tag{5}$$

$$\begin{aligned} \sum y_i x_{i,2} &= \sum \left(\sum_{j=2}^p \beta_j x_{i,j} x_{i,2} \right) + \sum x_{i,2} (u_i - \bar{u}) \\ \sum y_i x_{i,3} &= \sum \left(\sum_{j=2}^p \beta_j x_{i,j} x_{i,3} \right) + \sum x_{i,3} (u_i - \bar{u}) \\ &\vdots \\ \sum y_i x_{i,p-1} &= \sum \left(\sum_{j=2}^p \beta_j x_{i,j} x_{i,p-1} \right) + \sum x_{i,p-1} (u_i - \bar{u}) \\ \sum y_i x_{i,p} &= \sum \left(\sum_{j=2}^p \beta_j x_{i,j} x_{i,p} \right) + \sum x_{i,p} (u_i - \bar{u}) \end{aligned}$$

Now, we multiply both sides of the equation by $\frac{1}{x_{i,2}^2}$

$$\begin{aligned} \frac{\sum y_i x_{i,2}}{\sum x_{i,2}^2} &= \beta_2 \frac{\sum x_{i,2}^2}{\sum x_{i,2}^2} + \beta_3 \frac{\sum x_{i,3} x_{i,2}}{\sum x_{i,2}^2} + \frac{\sum x_{i,2} (u_i - \bar{u})}{\sum x_{i,2}^2} \\ \frac{\sum y_i x_{i,2}}{\sum x_{i,2}^2} &= \frac{\beta_2 \sum x_{i,2}^2}{\sum x_{i,2}^2} + \frac{\beta_3 \sum x_{i,3} x_{i,2}}{\sum x_{i,2}^2} + \dots + \frac{\beta_{p-1} \sum x_{i,p} x_{i,2}}{\sum x_{i,2}^2} + \frac{\beta_p \sum x_{i,p} x_{i,2}}{\sum x_{i,2}^2} + \frac{\sum x_{i,2} (u_i - \bar{u})}{\sum x_{i,2}^2} \\ \frac{\sum y_i x_{i,3}}{\sum x_{i,3}^2} &= \frac{\beta_2 \sum x_{i,2} x_{i,3}}{\sum x_{i,3}^2} + \frac{\beta_3 \sum x_{i,3}^2}{\sum x_{i,3}^2} + \dots + \frac{\beta_{p-1} \sum x_{i,p-1} x_{i,3}}{\sum x_{i,3}^2} + \frac{\beta_p \sum x_{i,p} x_{i,3}}{\sum x_{i,3}^2} + \frac{\sum x_{i,3} (u_i - \bar{u})}{\sum x_{i,3}^2} \\ \frac{\sum y_i x_{i,p-1}}{\sum x_{i,p-1}^2} &= \frac{\beta_2 \sum x_{i,2} x_{i,p-1}}{\sum x_{i,p-1}^2} + \frac{\beta_3 \sum x_{i,3} x_{i,p-1}}{\sum x_{i,p-1}^2} + \dots + \frac{\beta_{p-1} \sum x_{i,p-1}^2}{\sum x_{i,p-1}^2} + \frac{\beta_p \sum x_{i,p} x_{i,p-1}}{\sum x_{i,p-1}^2} + \frac{\sum x_{i,p-1} (u_i - \bar{u})}{\sum x_{i,p-1}^2} \\ \frac{\sum y_i x_{i,p}}{\sum x_{i,p}^2} &= \frac{\beta_2 \sum x_{i,2} x_{i,p}}{\sum x_{i,p}^2} + \frac{\beta_3 \sum x_{i,3} x_{i,p}}{\sum x_{i,p}^2} + \dots + \frac{\beta_{p-1} \sum x_{i,p-1} x_{i,p}}{\sum x_{i,p}^2} + \frac{\beta_p \sum x_{i,p}^2}{\sum x_{i,p}^2} + \frac{\sum x_{i,p} (u_i - \bar{u})}{\sum x_{i,p}^2} \end{aligned} \tag{6}$$

$$\begin{aligned} \frac{\sum y_i x_{i,2}}{\sum x_{i,2}^2} &= \sum \left(\frac{\sum_{j=2}^p \beta_j x_{i,j} x_{i,2}}{\sum x_{i,2}^2} \right) + \frac{\sum x_{i,2} (u_i - \bar{u})}{\sum x_{i,2}^2} \\ \frac{\sum y_i x_{i,3}}{\sum x_{i,3}^2} &= \sum \left(\frac{\sum_{j=2}^p \beta_j x_{i,j} x_{i,3}}{\sum x_{i,3}^2} \right) + \frac{\sum x_{i,3} (u_i - \bar{u})}{\sum x_{i,3}^2} \\ &\vdots \\ \frac{\sum y_i x_{i,p-1}}{\sum x_{i,p-1}^2} &= \sum \left(\frac{\sum_{j=2}^p \beta_j x_{i,j} x_{i,p-1}}{\sum x_{i,p-1}^2} \right) + \frac{\sum x_{i,p-1} (u_i - \bar{u})}{\sum x_{i,p-1}^2} \\ \frac{\sum y_i x_{i,p}}{\sum x_{i,p}^2} &= \sum \left(\frac{\sum_{j=2}^p \beta_j x_{i,j} x_{i,p}}{\sum x_{i,p}^2} \right) + \frac{\sum x_{i,p} (u_i - \bar{u})}{\sum x_{i,p}^2} \end{aligned}$$

Now, we have the familiar equations:

$$\begin{aligned}
\frac{\sum y_i x_{i,2}}{\sum x_{i,2}^2} &= \beta_2 \underbrace{\frac{\sum x_{i,2}^2}{\sum x_{i,2}^2}}_1 + \beta_3 \frac{\sum x_{i,3} x_{i,2}}{\sum x_{i,2}^2} + \frac{\sum x_{i,2} (u_i - \bar{u})}{\sum x_{i,2}^2} \\
\frac{\sum y_i x_{i,2}}{\sum x_{i,2}^2} &= \beta_2 \frac{\sum x_{i,2}^2}{\sum x_{i,2}^2} + \beta_3 \frac{\sum x_{i,3} x_{i,2}}{\sum x_{i,2}^2} + \dots + \beta_{p-1} \frac{\sum x_{i,p-1} x_{i,2}}{\sum x_{i,2}^2} + \beta_p \frac{\sum x_{i,p} x_{i,2}}{\sum x_{i,2}^2} + \frac{\sum x_{i,2} (u_i - \bar{u})}{\sum x_{i,2}^2} \\
\frac{\sum y_i x_{i,2}}{\sum x_{i,p}^2} &= \beta_2 \frac{\sum x_{i,2} x_{i,p}}{\sum x_{i,p}^2} + \beta_3 \frac{\sum x_{i,3} x_{i,p}}{\sum x_{i,p}^2} + \dots + \beta_{p-1} \frac{\sum x_{i,p-1} x_{i,p}}{\sum x_{i,p}^2} + \beta_p \frac{\sum x_{i,p} x_{i,p}}{\sum x_{i,p}^2} + \frac{\sum x_{i,2} (u_i - \bar{u})}{\sum x_{i,2}^2} \tag{7}
\end{aligned}$$

$$b_{12} = \beta_2 + \beta_3 b_{32} + \frac{\sum x_{i,2} (u_i - \bar{u})}{\sum x_{i,2}^2} \tag{8}$$

Now, let's take the expectation on both sides of the equation

$$E(b_{12}) = E(\beta_2) + E(\beta_3 b_{32}) + E\left(\underbrace{\frac{\sum x_{i,2} (u_i - \bar{u})}{\sum x_{i,2}^2}}_0\right) \tag{9}$$

From the properties of the expectation operator and the assumptions of classical linear regression, $E(b_{12}) = b_{12}$, $E(\beta_2) = \beta_2$, $E(\beta_3 b_{32}) = \beta_3 b_{32}$, and $E(x_{i,2} u_i) = 0$, $\bar{u} = 0$. Then, omitted variable bias can be shown as follows:

$$b_{12} = \beta_2 + \beta_3 b_{32} \tag{10}$$

$$\begin{aligned}
b_{12} &= \beta_2 + \beta_3 b_{32} + \beta_4 b_{42} + \dots + \beta_{p-1} b_{p-1,2} + \beta_p b_{p,2} \\
b_{13} &= \beta_2 b_{23} + \beta_3 + \beta_4 b_{43} + \dots + \beta_{p-1} b_{p-1,3} + \beta_p b_{p,3} \\
&\vdots \\
b_{1,p-1} &= \beta_2 b_{2,p-1} + \beta_3 b_{3,p-1} + \beta_4 b_{4,p-1} + \dots + \beta_{p-1} + \beta_p b_{p,p-1} \\
b_{1,p} &= \beta_2 b_{2,p} + \beta_3 b_{3,p} + \beta_4 b_{4,p} + \dots + \beta_{p-1} b_{p-1,p} + \beta_p
\end{aligned}$$

This derivation can also be represented with matrix algebra (see appendix A for the matrix algebra notations).

2.2 Total differentials and derivative

The partial derivative definition requires that there is no functional dependence between the independent variables, such that any independent variable can change without affecting the values of the other independent variables. If the independent variable is related to other independent variables, we cannot directly take the partial derivative and transfer the change in the related independent variable to the dependent variable. In this case, it will indirectly affect the dependent variable through the other related independent variables. Therefore, an operator should be developed that can handle this situation more properly than the partial derivative does. In these cases, it is necessary to resort to total derivatives instead of partial derivatives. In order to first understand the total differentiation process, originating from the concept of total derivative, it is necessary to understand the concept of differential [2]. When we say total derivative, we discuss the case where dy or dx are treated as separate. However, we characterize these isolated structures as dy / dx derivatives. We can extend this structure to more than one variable. In this case, the concept of differential turns into the concept of total derivative.

Given a function such as $y = f(x)$, the difference division $\Delta y/\Delta x$ represents the rate of change of y with respect to x . Since $\Delta y = (\Delta y/\Delta x)\Delta x$ is valid, given the rate of change of $\Delta y/\Delta x$ and the change in x , the magnitude of Δy can be found. When Δx is very small, Δy will also be very small, and the difference quotient $\Delta y/\Delta x$ will turn into the derivative of dy/dx . In this case, small changes in x and y will transform Δy into dy and Δx into dx . Now, we can represent the identity with the following equation:

$$dy = \left(\frac{dy}{dx}\right)dx \text{ or } dy = f'(x)dx \quad (11)$$

The symbols dy and dx are called the differential of y and x , respectively. If we divide the identity (11) by dx , we obtain:

$$\frac{(dy)}{(dx)} = \left(\frac{dy}{dx}\right)\frac{dx}{dx} \text{ or } \frac{dy}{dx} = f'(x)\frac{dx}{dx} \quad (12)$$

$$\frac{(dy)}{(dx)} = \left(\frac{dy}{dx}\right) \text{ or } \frac{dy}{dx} = f'(x)$$

In this way, it is seen that the division of differential dy and dx can be interpreted as a derivative. Once the derivative of a function is given, the differential dy is immediately obtained. This result can be used to calculate the change in y caused by a change in x . However, it should be noted that the dy and dx differentials are only valid for very small changes. The concept of differential can easily be extended to include functions with two or more independent variables $y = f(x_2, x_3)$ or $Y_t = f(X_{t,2}, X_{t,3})$. Here, the variables have been denoted as in the multiple regression estimation to be consistent with the Gujarati and Porter [1] (see p.185). The total change in Y_t due to small changes in $X_{t,2}$, and $X_{t,3}$ is represented by differentials and partial derivatives as follows:

$$dY_t = \frac{\partial Y_t}{\partial X_{t,2}}dX_{t,2} + \frac{\partial Y_t}{\partial X_{t,3}}dX_{t,3} \quad (13)$$

$$dY_t = f'(X_{t,2})dX_{t,2} + f'(X_{t,3})dX_{t,3} \quad (14)$$

The dY_t differential is the sum of the change stemming from two sources, changes in $X_{t,2}$ and $X_{t,3}$. Finding such a total differential is called as a total derivative. For example, if any $X_{t,3}$ is constant, then $dX_{t,3}$ will be zero. Then, we have

$$dY_t = \frac{\partial Y_t}{\partial X_{t,2}}dX_{t,2} \quad (15)$$

If we divide both sides of equation (15) by $dX_{t,2}$:

$$\frac{dY_t}{dX_{t,2}} = \frac{\partial Y_t}{\partial X_{t,2}} \frac{dX_{t,2}}{dX_{t,2}} \quad (16)$$

In this case, we can also show that the total differential and the partial derivative are equal. We can generalize this into n variables as follows:

$$dY_t = f'(X_{t,2})dX_{t,2} + f'(X_{t,3})dX_{t,3} + \dots + f'(X_{t,n})dX_{t,n} = \sum_{i=1}^n f'(X_{t,i})dX_{t,i} \quad (17)$$

Since the concept of differential has explained, the next step is to show the rate of change in Y_t with respect to, when $X_{t,2}$ and $X_{t,3}$ are connected or correlated. The methodology used in such cases is multivariate regression analysis. Both Gujarati and Porter [1] and Chiang and Wainwright [2] describe exactly the same relationship regarding to the subject in the context of multivariate regression. Chiang and Wainwright [2] describes it through a figure, which they refer as a channel map on page 190, while Gujarati and Porter [1] explains it through direct and indirect effects. The relationship is mathematically explained by Chiang and Wainwright [2] by the help of two functions $Y_t = f(X_{t,2}, X_{t,3})$ and $X_{t,3} = g(X_{t,2})$. They argue that $X_{t,2}$ can affect the dependent variable Y_t through two separate channels: (1) indirectly, through $X_{t,3}$ (via the function g and then with the f function), and (2) directly, via the function f . The direct effect can simply be represented by the partial derivative $fX_{t,2}$ while the indirect effect can only be expressed by a product of two derivatives. In other words, a total derivative, $dY_t = fX_{t,2}dX_{t,2} + fX_{t,3}dX_{t,3}$ is needed to express both effects together in this case. To get this total derivative, we use differentials and partial derivatives. If both sides of this total derivative are divided by the $dX_{t,2}$ differential:

$$\begin{aligned} \frac{dY_t}{dX_{t,2}} &= fX_{t,2} \frac{dX_{t,2}}{dX_{t,2}} + fX_{t,3} \frac{dX_{t,3}}{dX_{t,2}} \\ \frac{dY_t}{dX_{t,2}} &= fX_{t,2} + fX_{t,3} \frac{dX_{t,3}}{dX_{t,2}} \end{aligned} \quad (18)$$

Since the ratio of two differentials to each other can be interpreted as a derivative, the expression $dY_t / dX_{t,2}$ on the left-hand side will be a measure of the rate of change of Y_t relative to $X_{t,2}$. Moreover, from the above equation, we see the indirect effect of how $X_{t,2}$ affects Y_t through $X_{t,3}$.

$$\left(\frac{\partial Y_t}{\partial X_{t,3}} \right) \left(\frac{dX_{t,3}}{dX_{t,2}} \right)$$

We have demonstrated the indirect effect with partial derivative and differential operators. In addition, we see how to take the total derivative of Y_t with respect to $X_{t,2}$. We can next represent omitted variable bias by using the total derivative method.

3. A new method to explain omitted variable bias via total differential

We have shown how to take the total derivative of Y_t with respect to variable $X_{t,2}$. This total derivative when the variable $X_{t,3}$ is not included in the regression equation, causes a bias in the β_2 parameter. Now, we have come to the part where we show how omitted variable bias and total derivative are related to each other. For this purpose, we first take the total derivative of Y_t with respect to $X_{t,2}$, and then we show the multivariate regression equivalents of the total and partial derivatives. Unlike the partial derivative, a total derivative does not require $X_{t,3}$ to remain constant as $X_{t,2}$ changes, thus allows the predicted relationship.

$$\begin{aligned} \frac{dY_t}{dX_{t,2}} &= fX_{t,2} \frac{dX_{t,2}}{dX_{t,2}} + fX_{t,3} \frac{dX_{t,3}}{dX_{t,2}} \\ \frac{dY_t}{dX_{t,2}} &= \frac{\partial Y_t}{\partial X_{t,2}} \frac{dX_{t,2}}{dX_{t,2}} + \frac{\partial Y_t}{\partial X_{t,3}} \frac{dX_{t,3}}{dX_{t,2}} \\ \frac{dY_t}{dX_{t,2}} &= \frac{\partial Y_t}{\partial X_{t,2}} + \frac{\partial Y_t}{\partial X_{t,3}} \frac{dX_{t,3}}{dX_{t,2}} \\ \frac{dY_t}{dX_{t,2}} &= \frac{\partial Y_t}{\partial X_{t,2}} + \frac{\partial Y_t}{\partial X_{t,3}} \frac{dX_{t,3}}{dX_{t,2}} \\ \underbrace{\frac{dY_t}{dX_{t,2}}}_{\beta_2} &= \underbrace{\frac{\partial Y_t}{\partial X_{t,2}}}_{b_{12}} + \underbrace{\frac{\partial Y_t}{\partial X_{t,3}}}_{b_{32}} \underbrace{\frac{dX_{t,3}}{dX_{t,2}}}_{\beta_3} \end{aligned} \quad (19)$$

Finally, omitted variable bias can be expressed as:

$$b_{12} = \beta_2 + \beta_3 b_{32} \quad (20)$$

We complement this structure with Figure 1, which depicts a relationship between omitted variable bias and total derivative. The lower right panel of the figure illustrates the function $X_{t,3} = f(X_{t,2})$, showing the effect of $X_{t,2}$ on $X_{t,3}$. We see that when there is a change in $X_{t,2}$, there will be a change in $X_{t,3}$. By the help of a mirror graph containing the 45-degree line in the lower left panel, this change in $X_{t,3}$ can be reflected to the upper left 3D panel showing the indirect effect of $X_{t,2}$ on Y_t through $X_{t,3}$. The derivations and graphical explanations provided above show the relationship of the omitted variable bias with the total derivative. MATLAB code for Figure 1 is provided in Appendix B. We have provided a numerical example in Appendix C, which serves as a guide for readers who seek a practical illustration. It would ensure that the demonstration in Figure 1 is effectively interpreted and contextualized.

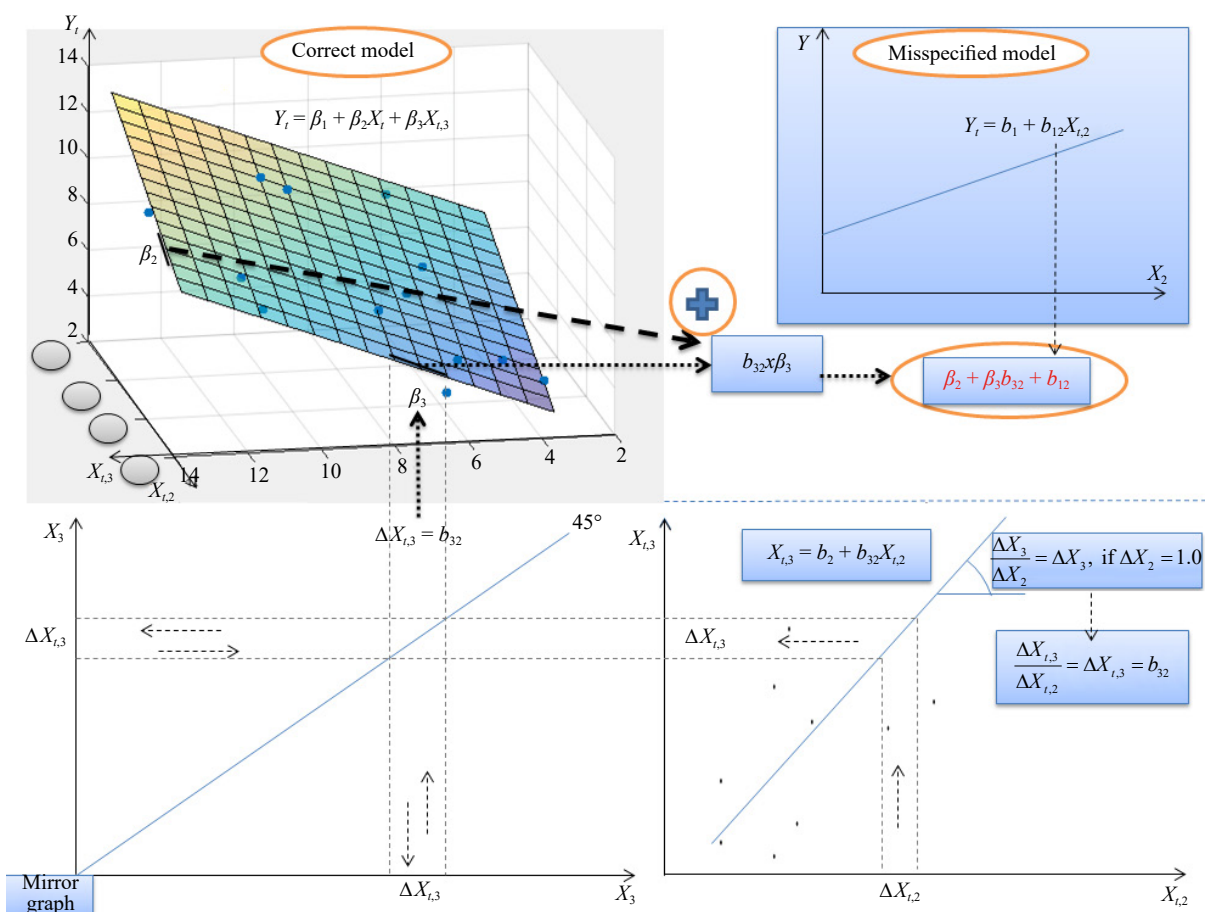


Figure 1. The flow chart of omitted variable bias via total derivative

4. Concluding remarks

In this study, we try to show how the omitted variable bias problem can be discussed within the context of the total derivative method. We argue that discussing the concepts of multivariate regression and the total derivative method together would bring the fields of statistics (or econometrics) and mathematics closer together. The application of the total derivative method to the omitted bias problem contributes to a better understanding of regression analysis and fills an important gap in the field. It may inspire similar explorations of other statistical or econometric issues. Furthermore, a

graphical representation of the method provided here may be constructive for mathematics and statistics or econometrics researchers in visualizing the concepts of total derivative and omitted variables.

This new approach emerged from the present study and could also be applied to other statistics or econometrics issues. Based on the understanding here, cases such as the estimation of regression including the correlation intersection set of two correlated variables alone or the model misspecifications that may arise from leaving out such variables can be examined with the help of the total differential method in future research. Similarly, model misspecifications stemming from a correlation between the error term and the dependent variables could also be explained by the total differential method. Multicollinearity, cross-section dependency, unit root methods including covariate, models including volatility and covariance, and similar issues in econometrics can be examined in more detail by using the total differential method (see Emirmahmutođlu [3], Hansen [4], Samuilik et al. [5], and Manickam et al. [6]).

Conflict of interest

The authors have no conflict of interest either wholly or partially in the content of the article.

References

- [1] Gujarati DN, Porter DC. *Basic econometrics*. 5th ed. New York: McGraw-Hill-Irwin; 2009.
- [2] Chiang A, Wainwright K. *Fundamental methods of mathematical economics*. 4th ed. New York: McGraw-Hill-Irwin; 2005.
- [3] Emirmahmutođlu F. Cross-section dependency and the effects of nonlinearity in panel unit testing. *Econometrics Letters*. 2014; 1(1): 30-36.
- [4] Hansen BE. Rethinking the univariate approach to unit root testing: Using covariates to increase power. *Econometric Theory*. 1995; 11(5): 1148-1171.
- [5] Samuilik I, Sadyrbaev F, Ogorelova D. Comparative analysis of models of gene and neural networks. *Contemporary Mathematics*. 2023; 4(2): 217-229. Available from: <https://doi.org/10.37256/cm.4220232404>.
- [6] Manickam A, Indrakala S, Kumar P. A novel mathematical study on the predictions of volatile price of gold using grey models. *Contemporary Mathematics*. 2023; 4(2): 270-285. Available from: <https://doi.org/10.37256/cm.4220232389>.
- [7] Greene W. *Econometric analysis*. 5th ed. Pearson Education; 2002.
- [8] Gujarati DN. *Basic econometrics*. International 3rd ed. İstanbul: McGraw-Hill-Literatür; 1995.

Appendix A

We have shown the omitted variable bias by using matrix algebra for the interested reader [7].

$$y = X\beta + u \quad (\text{A.1})$$

$$y = b_1 X_1 + \varepsilon \quad (\text{A.2})$$

$$b_1 = (X_1' X_1)^{-1} X_1' y \quad (\text{A.3})$$

$$b_1 = \beta_1 + (X_1' X_1)^{-1} X_1' X_2 \beta_2 + (X_1' X_1)^{-1} X_1' e \quad (\text{A.4})$$

$$E[b_1 | X] = \beta_1 + P_{12} \beta_2 \quad (\text{A.5})$$

$$P_{12} = (X_1' X_1)^{-1} X_1' X_2 \quad (\text{A.6})$$

Appendix B

MATLAB code for the Figure 1.

```
x = [14.9 1.7 0.0 10.9 0.0];
y = [11.3 9.1 23.7 12.8 2.9];
z = [5.32787E-17 2.93234E-16 2.09997E-16 5.45E-17 4.55E-16];
B = [x(:) y(:) ones(size(x(:))) \ z(:)];
xv = linspace(min(x), max(x), 10)';
yv = linspace(min(y), max(y), 10)';
[X,Y] = meshgrid(xv, yv);
Z = reshape([X(:), Y(:), ones(size(X(:)))] * B, numel(xv), []);
scatter3(x,y,z, 'filled')
hold on
mesh(X, Y, Z, 'FaceAlpha', 0.5)
hold off
view(-120, 35)
title(sprintf('Z = %+.3E\cdotX %+.3E\cdotY %+.3E', B))
```

Appendix C: Application of the method

For illustration purposes, let us take the example from Gujarati [8] (see p.203). The yearly data for the time period 1970-1978 for the United States is considered to estimate a Philips curve relationship. We extend the implementation period and repeat the same analysis by incorporating the years 2010-2022 for the US. We collect the data from from the Federal Reserve Bank of St. Louis. This inclusion allowed us to incorporate the dynamics of the Philips curve relationship into our applied study, assessing whether it continues to hold. We use the t subscript since the variables are time series.

Table B1. Actual inflation rate Y_t (%), unemployment rate $X_{t,2}$ (%), expected inflation rate $X_{t,3}$ (%), USA, 1970-1982/2010-2022

Year	Y_t^*	$X_{t,2}$	$X_{t,3}$
1970	5.92	4.90	4.78
1971	4.30	5.90	3.84
1972	3.30	5.60	3.13
1973	6.23	4.90	3.44
1974	10.97	5.60	6.84
1975	9.14	8.50	9.47
1976	5.77	7.70	6.51
1977	6.45	7.10	5.92
1978	7.60	6.10	6.08
1979	11.47	5.80	8.09
1980	13.46	7.10	10.01
1981	10.24	7.60	10.81
1982	5.99	9.70	8.00
2010	1.64	9.30	1.49
2011	3.16	8.96	1.83
2012	2.07	8.08	1.53
2013	1.46	7.36	1.47
2014	1.62	6.16	1.69
2015	0.12	5.28	1.39
2016	1.26	4.88	1.60
2017	2.13	4.36	1.84
2018	2.44	3.89	2.01
2019	1.81	3.68	1.75
2020	1.23	8.09	1.16
2021	4.70	5.37	2.04
2022	8.00	3.64	3.25

Notes: The data set for the years 1978-1982 is sourced from Gujarati [8] where Y_t and $X_{t,2}$ data are collected from a data source of Business Statistics, 1982, from various pages of the US Department of Commerce Bureau of Economic Analysis; $X_{t,3}$ data are collected from the various pages of the Economic Review, Federal Reserve Bank of Richmond. The data for the years 2010-2022 is authors' own calculations and is sourced from the Federal Reserve Bank of St. Louis.

* shows the percentage change in the Consumer Price Index.

Gujarati [8] used this example to illustrate omitted variable bias:

$$Y_t = \beta_1 + \beta_2 X_{t,2} + \beta_3 X_{t,3} + u_{t,1} \tag{B1}$$

Y_t is the actual inflation rate in period t , $X_{t,2}$ is the unemployment rate in period t , and $X_{t,3}$ is the expected inflation rate in period t . Extended Phillips curve with expectations will be obtained by the above regression analysis. According to macroeconomic theory, the effect of β_2 is negative. In addition, the effect of β_3 is positive and $\beta_3 = 1.0$.

$$Y_t = 3.045 - 0.442 X_{t,2} + 1.550 X_{t,3} \tag{B2}$$

(2.587) (-2.380) (10.788)

$$R^2 = 0.914$$

```

MATLAB code
num = xlsread('data.xlsx') % read data from Excel
y = num(:,1) % Column Vector
x1 = num(:,2) % Column Vector
x2 = num(:,3) % Column Vector
x3 = num(:,4) % Column Vector
x = [x1 x2 x3] % create 3*3 matrix
B = inv(x'*x)*x'*y % use OLS formula
k=3; % use this for degrees of freedom and calculate the variance
error=y-x*B %obtain residuals for t-value
A=inv(x'*x); % obtain variance covariance matrix of error term
SSR=(error'*error); % obtain residual sum of squares
sigma2=SSR/(length(y)-k); % obtain variance
tvalue=B./sqrt(sigma2*diag(A)) % obtain t-value for the estimates

```

The numbers in parentheses give the t values. If $X_{t,2}$ and $X_{t,3}$ are held constant during the sample period, the average actual inflation will be around 3.04 percent. The partial regression coefficient of -0.442 means that, holding $X_{t,3}$ (expected inflation rate) constant, the actual inflation rate will increase (decrease) by an average of 0.44 percent for each unit (here one percentage point) decrease (increase) in the unemployment rate over the period 1970-1982/2010-2022. Likewise, the coefficient of 1.15 shows that, keeping the unemployment rate constant, for every one percent increase in the expected inflation rate, the actual inflation rate will increase by 1.15 percent on average. In the context of preliminary expectations, both variables bear the expected signs. According to the classical linear regression model assumptions, the regression model used in the analysis is correctly established. So, there is no model specification error or deviation. Now, let's examine the situation where we did not include the $X_{t,3}$ variable, that is, the expected inflation rate, into the model by creating a model-specification deviation.

$$Y_t = b_1 + b_{12}X_{t,2} + u_{t,2} \tag{B3}$$

In this case, we will try to show that the variable $X_{t,2}$ contains the effects of the variable $X_{t,3}$, which was not included in the model, that is, the variable b_{32} carries the effects of the variable $X_{t,3}$. It will be shown that the $\beta_2 b_{32}$ parameter makes a biased estimate due to the effects of $X_{t,3}$. Therefore, an important question arises: how to calculate the amount of this bias? We have already shown this bias in equations (9) and (10).

$$E(b_{12}) = \beta_2 + \beta_3 b_{32} \tag{B4}$$

So, we know that $b_{12} \neq \beta_2$ is not equal to each other. Here, the numerical amount of this bias will be shown within the sample framework in question. The estimation of equation (B3) according to the sample is as follows:

$$Y_t = 4.183 + 0.143 X_{t,2} \tag{B5}$$

(1.480) (0.334)

$$r^2 = 0.068$$

MATLAB code
<pre> num = xlsread('data.xlsx') y =num(:,1) x1 = num(:,2) x2 = num(:,3) x =[x1 x2] B = inv(x'*x)*x'*y k=2; error=y-x*B A=inv(x'*x); SSR=(error'*error); sigma2=SSR/(length(y)-k); tvalue=B./sqrt(sigma2*diag(A)) </pre>

As can be seen, $b_{12} = 0.143$ has a positive sign and is statistically insignificant. However, from equation (B2), it is estimated as $b_{12} \neq \beta_2 = -0.442$, where it has a negative sign and is statistically significant. We have obtained a biased estimation since the effects of the $X_{t,3}$ variable are reflected in the b_{12} parameter estimation. Now, in order to calculate this bias, namely β_2 , let us estimate the regression equation for the following variables $X_{t,3}$ and $X_{t,2}$ to obtain b_{32} . β_2 and β_3 parameters are already obtained in equation B2.

$$X_{t,3} = \underset{(0.440)}{0.988} + \underset{(1.496)}{0.508} X_{t,2} \tag{B6}$$

$$r^2 = 0.292$$

MATLAB code
<pre> num = xlsread('data.xlsx') y =num(:,4) x1 = num(:,2) x2 = num(:,3) x =[x1 x2] B = inv(x'*x)*x'*y k=2; error=y-x*B A=inv(x'*x); SSR=(error'*error); sigma2=SSR/(length(y)-k); tvalue=B./sqrt(sigma2*diag(A)) </pre>

As can be seen from the regression equation (B6), b_{32} is estimated as $b_{32} = 0.508$. The b_{12} omitted variable bias can now be calculated.

$$b_{12} = \hat{\beta}_2 + \hat{\beta}_3 b_{32} = -0.442 + 1.150 \times 0.508 = 0.143 \tag{B7}$$

As seen from equation (B7), the bias in $\beta_3 b_{32}$ value is added to the β_2 parameter, and the b_{12} parameter is found. Indirect and direct effects are clearly explained.

For the total differentiation we can use the below MATLAB codes for computations:

MATLAB code

```
syms y x2 x3
y=3.045-0.442*x2+1.115*x3;
x3=0.988+0.508*x2;
B2 = diff(y(x2,x3),x2);
B3 = diff(y(x2,x3),x3);
b32 = diff(x3(x2),x2);
b12 = B2+B3*b32 % Total Differentiation for x2
```