Introduction

The quantitative study of phenomena seeks to estimate the effect of a variable (or more) on another variable. When one (or more) variables are left out, the estimate can be erroneous. Even increasing the sample will not alleviate this problem (Jargowsky, 2005). Quantitative social sciences are especially susceptible to omitted variable bias since social phenomena tend to have multivariate relationships and not bivariate. Most of the time assuming a bivariate relationship tends to be incorrect. It is not uncommon for researchers to assume bivariate relationships due to factors like graphical representations of data which are usually effective in demonstrating bivariate relationships. (Jargowsky, 2005). When a variable which has an effect on the independent and dependent variable is omitted, this is called the omitted variable bias (Wilms et al, 2021).

Our work seeks to do three things. First, it provides context and description of omitted variable bias and the ways in which it can be alleviated. Second, it describes a brief simulation study and presents the result of said study. Finally, in the third section, the major results of this project are discussed.

Part I: Background, theory, and methodology

Background

Linear regression is used to examine causality in research. A linear regression quantifies a linear relationship between two variables. The variable y_i is explained by the variable x_i (with slope β_1) and intercept β_0 . The term e_i is the error term.

$$y_i = \beta_0 + \beta_1 x_i + e_i,$$

Linear regression is calculated by using Ordinary least squares estimators which minimizes the sum of e_i by choosing β_0 and β_1.

In order to use linear regression certain assumptions need to hold, if these assumptions do not hold then there is the risk of producing a mispsecified model which does not reflect reality. One of the assumptions of linear regression is the exogeneity assumption which is violated by omitted variable bias(Wilms et al, 2021). That is to say, variable x_i needs to be independent from e_i. This assumption implies that the covariance between e_i and x_i. If this assumption is violated then the coefficient will be over/underestimated and can be inconsistent, and it could increase with sample size.

While there are many ways in which the exogeneity assumption can be violated, one of the most common ones is the omitted variable bias. The omitted variable bias occurs when a variable that affects the independent variable(s) (x_i) and dependent variable (y_i) is omitted.

Endogeneity Bias

We suppose the correct model is:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 q_i + \epsilon_i,$$

Where,

$$x_i = \lambda_1 q_i + \mu_i,$$

Because of measurement inaccuracy, we cannot observe the q_i. We try to find the relationship between x_i and y_i by a misspecified model where x_i and ϵ_i are linked:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\begin{tikzcd}
x_i \arrow[rr] \arrow[dr, leftrightarrow] & & y_i \\
& \epsilon_i \arrow[ur] &
\end{tikzcd}$$

By OLS method, we find the estimator of β_1 in the misspecified. It has 2 parts, one is the effect of changing x_i on y_i; the other is the change in x_i produced by ϵ_i, and hence the change in y_i.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Now, let us try to get a more precisely estimator of β_1 by plug the $y_i = \beta_0 + \beta_1 x_i + \beta_2 q_i + \epsilon_i$ in the upper equation, then, we get:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + \beta_2 q_i + \epsilon_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

For the sake of simplicity, we'll suppose that x, y, and q all have the same mean 0 and that x is a fixed variable rather than a random variable. The upper equation becomes:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i)(\beta_0 + \beta_1 x_i + \beta_2 q_i + \epsilon_i)}{\sum_{i=1}^{n} {x_i}^2}$$

$$= \frac{\sum_{i=1}^{n}(\beta_0 x_i + \beta_1 {x_i}^2 + \beta_2 x_i q_i + x_i \epsilon_i)}{\sum_{i=1}^{n} {x_i}^2}$$

$$= \frac{\beta_0 \sum_{i=1}^{n} x_i + \beta_1 {x_i}^2 + \beta_2 x_i q_i + x_i \epsilon_i)}{\sum_{i=1}^{n} {x_i}^2}$$

$$= \frac{\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} {x_i}^2 + \beta_2 \sum_{i=1}^{n} x_i q_i + \sum_{i=1}^{n} x_i \epsilon_i}{\sum_{i=1}^{n} {x_i}^2}$$

$$= \frac{\beta_1 \sum_{i=1}^{n} {x_i}^2 + \beta_2 \sum_{i=1}^{n} x_i q_i + \sum_{i=1}^{n} x_i \epsilon_i}{\sum_{i=1}^{n} {x_i}^2}$$

$$= \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} x_i q_i}{\sum_{i=1}^{n} {x_i}^2} + \frac{\sum_{i=1}^{n} x_i \epsilon_i}{\sum_{i=1}^{n} {x_i}^2}$$

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 E(\frac{\sum_{i=1}^{n} x_i q_i}{\sum_{i=1}^{n} {x_i}^2}) + \frac{\sum_{i=1}^{n} x_i E(\epsilon_i)}{\sum_{i=1}^{n} {x_i}^2}$$

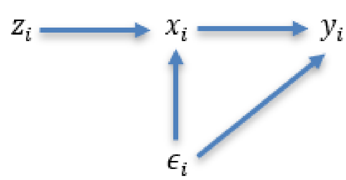
$$= \beta_1 + \beta_2 E(\frac{\sum_{i=1}^{n} x_i q_i}{\sum_{i=1}^{n} {x_i}^2})$$

Based on the upper equation, we can see if there exists endogeneity, the $\hat{\beta}_1$ we get from OLS is not an unbiased estimator. Apparently, $\beta_2 E(\frac{\sum_{i=1}^{n} x_i q_i}{\sum_{i=1}^{n} x_i^2})$ is the bias. When $\beta_2 E(\frac{\sum_{i=1}^{n} x_i q_i}{\sum_{i=1}^{n} x_i^2})$ is a positive value, we obtain a positive bias; when $\beta_2 E(\frac{\sum_{i=1}^{n} x_i q_i}{\sum_{i=1}^{n} x_i^2})$ is a negative value, we obtain a negative bias.

Statistical methods to correct endogeneity from Omitted Variable Bias

There are multiple ways to address endogeneity from omitted variable bias, while there are theoretical and experimental approaches to this problem, this paper will focus on statistical methods to alleviate said issue.

One of the most widely used approaches to solve endogeneity issues such as omitted variable bias is to use the two-stage least squares estimator (*2SLS*) and instrument variable approach. By introducing an instrumental variable, we hope to isolate the effect of x on y.



This approach consists of two steps. First the effect of an instrument variable (z_i) on x_i is calculated. The meaning of this step is to divide x_i into two parts, a part determined by z_i and a part independent of z_i, which are orthogonal to each other. Because the definition of instrumental variables is related to x_i rather than ϵ_i, it follows that the two parts of x_i separated by z_i are also unrelated to ϵ_i. The function of z_i is that a change in z_i causes a change in x_i which in turn causes a change in y_i. This effect is purely the effect of x_i on y_i. The effects of $z_i$on y_i are all generated through x_i and do not include ϵ_i because z_i and ϵ_i are uncorrelated, so that the coefficients β_1 of y_i can be estimated in this way without bias.

$$x_i = \pi_0 + \pi_1 z_i + v_i$$

The second step consists of using this estimation $\hat{x}_i$ to predict y, using these two steps makes the exogeneity assumption consistent since $\hat{x}_i$ is exogenous.

$$\hat{x}_i = \hat{\pi}_0 + \hat{\pi}_1 \, z_i$$

$$\hat{y}_i = \beta_0 + \beta_{IV} \hat{x}_i + e_i$$

Part II: Empirical studies

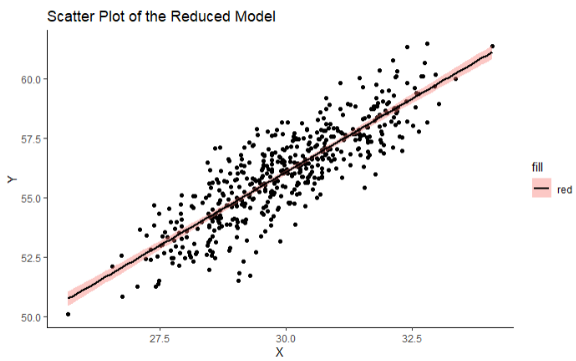
The goal of our simulation analysis is to use simulated data to examine how the model with the omitted variable affects the results of the linear regression estimation in comparison to the true model. Furthermore, the article approaches the use of an instrumental variable in order to alleviate the effects of the omitted variable and compares these to the results of the correctly specified model. A comparison of these models is then conducted. Based on a blog post by Simmering (2014), a simulation was conducted, using *R*.

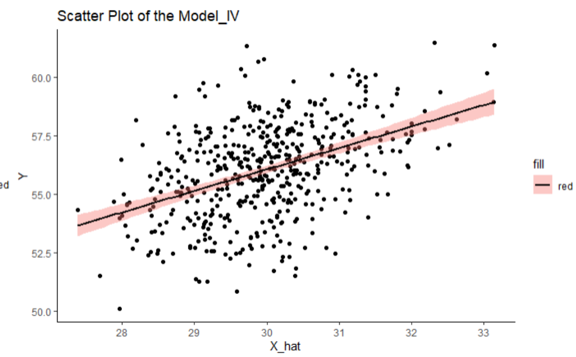
The package *MASS* (Ripley et al 2013) was used to create normally distributed data. We used MASS to simulated a set of data with 2 columns which are correlated. The variable *X* is defined as the normally distributed variable, *Z* in addition to one column of the normally distributed data. The variable *Q* is defined as another simulated data with normal distribution. The variable *X* has a correlation with *Q* and *Z*, however *Q* and *Z* do not have a correlation. *Y* is defined as

Y=1+X+Q+e.*

**e* is a normally distributed variable which accounts for the error

Using simulated data, three models were created to investigate the effects of omitted variables and the corrective effects of two-stage least squares and instrumental variables methods for correctly estimating the parameters.





As shown in Table 1, *Model_True* shows the result of the correct model, β_1 and coefficient of $Q(\beta_2)$ are all very close to 1, the *T* statistics for β_1 is 64.596 which is very large, R^2 is 0.9462 which shows that we made a very good fit here; *Model_Reduced* shows the results of the model with omitted *Q*, the coefficient of *X* is 1.23089 which is apparently biased with the true coefficient and the *T* statistics gets smaller, the R^2 is 0.7419 which is smaller; *Model_IV* shows the regression results using the two-stage least squares and instrumental variables methods, β_1 is 1.02130 which is very close to the true value, the R^2 is 0.2612, it's small but probably because we haven't conclude the whole parameter in this regression. Based on the results, the methods of two-stage least square and instrumental variable can definitely help us obtain a much precise estimator of β_1.

Table 1 Results of Different Regression Methods

	Model_True	*Model_Reduced*	*Model_IV*
β_1	*1.01368****	*1.23089****	*1.02130****
	(64.596)	*(37.84)*	*(13.27)*
β_2	*1.00056****	-	-
	(43.416)	-	-
β_0	*0.56740*	*19.11760****	*25.41994****
	(0.917)	*(19.52)*	*(10.98)*
Methods	*Y=f(X, Q)*	*Y=f(X)*	$Y=f(\hat{X})$
Obs	*500*	*500*	*500*
F-value	*4367****	*1432****	*176.1****
R^2	*0.9462*	*0.7419*	*0.2612*

In order to explore whether the bias caused by omitted variables would be cut with increasing sample size, we used simulated regression models with different sample sizes, and their regression results can be seen in the Table 2. According to the experimental results, we can see that when the sample size is very small, n=50, we do not get the correct estimator of β_1 even with the correct regression model; and when the sample size is large enough, n=500 and n=5000, the bias caused by the omitted variables cannot be reduced. For example, the estimator of β_1 in the reduced model when n=500 is 1.23089 and when n=5000 is 1.24878, the bias even becomes bigger.

Table 2 Results of Different Regression Methods with Different Sample Size

	n=50			n=500			n=5000		
	Model_True	*Model_Reduced*	*Model_IV*	*Model_True*	*Model_Reduced*	*Model_IV*	*Model_True*	*Model_Reduced*	*Model_IV*
β_1	*0.95102****	*1.3236****	*0.6226**	*1.01368****	*1.23089****	*1.0213****	*1.00246****	*1.24878****	*0.98671****
	(16.709)	*(9.819)*	*(2.154)*	*(64.596)*	*(37.84)*	*(13.27)*	*(188.581)*	*(117.48)*	*(62.36)*
β_2	*1.09350****	-	-	*1.00056****	-	-	*0.99283****	-	-
	(16.507)	-	-	*(43.416)*	-	-	*(133.193)*	-	-
β_0	*0.01653*	*16.0701****	*37.295****	*0.56740*	*19.11760****	*25.420****	*1.11334****	*18.54657****	*26.41589****
	(0.009)	*(3.933)*	*(4.259)*	*(0.917)*	*(19.52)*	*(10.98)*	*(5.596)*	*(58.04)*	*(55.57)*
Methods	$Y=f(X, Q)$	$Y=f(X)$	$Y=f(\hat{X})$	$Y=f(X, Q)$	$Y=f(X)$	$Y=f(\hat{X})$	$Y=f(X, Q)$	$Y=f(X)$	$Y=f(\hat{X})$
R^2	*0.9511*	*0.6676*	*0.4804*	*0.9462*	*0.7419*	*0.2612*	*0.9416*	*0.7341*	*0.7018*

Part III: Discussion

The omitted variable bias has negative repercussions on the validity of research which seeks to estimate the causal relationship between variables. Through the simulation conducted in this project we showed how omitting a variable can affect the results of a linear regression, both giving a larger or smaller effect than the actual one. Furthermore, we showed how using an instrumental variable can help alleviate the effects of omitting a variable.

With regards to future research, omitted variable bias should be simulated with more variables in order to examine the effect of omitting a variable on the estimates of a linear regression with multiple variables.

However, the key to the instrumental variables approach is the selection of a valid instrumental variable. Due to the difficulties in the selection of instrumental variables, the instrumental variables approach itself suffers from two shortcomings: One is that the instrumental variable estimates are somewhat arbitrary because the instrumental variable is not unique; Second, since the error term is practically unobservable, it is in fact difficult to find variables that are strictly independent of the error term and highly correlated with the random explanatory variables they replace.

References

Jargowsky, P. A. (2005). Omitted variable bias. Encyclopedia of social measurement, 2, 919-924
.

Wilms, R., Mäthner, E., Winnen, L., & Lanwehr, R. (2021). Omitted variable bias: A threat to estimating causal relationships. Methods in Psychology, 100075.

Simmering, J. (2014, January 10). Instrumental variables simulation: R-bloggers. R. Retrieved December 10, 2021, from https://www.r-bloggers.com/2014/01/instrumental-variables-simulation/.

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., & Ripley, M. B. (2013). Package ‘mass’. Cran r, 538, 113-120.