DOI:10.2478/bile-2025-0009

Biometrical Letters
Vol. 62 (2025), No. 2, 141-143

A new accurate regression: Overcoming the specification error

Moawia Alghalith

Department of Economics, University of the West Indies at St. Augustine, Trinidad and Tobago, e-mail: malghalith@gmail.com

Summary

We develop a simple method that virtually eliminates the regression specification error and spurious relationships. Furthermore, we introduce a strong (real) causality test.

Key words: spurious regression, causality, specification error, OLS

1. Introduction

Identifying the omitted variables in a regression model has long been a significant obstacle in regression analysis. Other specification issues include irrelevant variables and incorrect functional forms. The traditional regression methods, in general, do not sufficiently address these problems. Existing approaches to addressing specification errors offer only partial and limited solutions (see, for example, Golden et al., 2016; Maddala and Lahiri, 2009).

Model misspecification can yield spurious regressions, especially when confounding variables are omitted (see Phillips, 1986; Asteriou and Hall, 2011; Alghalith, 2018). Such omissions also diminish the reliability of standard causality tests, including Granger causality and its variants (Granger, 1969). We emphasize that Granger's causality is not a real causality – it is a correlation test. Other causality tests have similar shortcomings and can be difficult to implement (see Shiffrin, 2016; Varian, 2016).

In this paper, we present a simple method that addresses the problem of model misspecification. Our proposed approach virtually eliminates the specification error. Moreover, the estimated parameters are extremely close to their true value, since the error is very small. A strong (real) causality test is also introduced.

2. The method

Choose any insignificant explanatory variable x_1 that is not expected to be correlated with the dependent variable y (see Alghalith, 2019) and run the following regression:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon, \tag{1}$$

where β is the parameter to be estimated, and ε is the error; so that the true model is

$$y = \beta_0 + \sum_{i=2}^{n} \beta_i x_i + \varepsilon_1, i \neq 1. \tag{2}$$

Clearly, $\varepsilon = \sum\limits_{i=2}^{n} \beta_i x_i + \varepsilon_1$.

Now, assume x_2 is the variable of interest. Use the residuals from (1) $\hat{\varepsilon}$ to run the following regression:

$$\hat{\varepsilon} = \bar{\beta}_0 + \bar{\beta}_2 x_2 + \varepsilon_2. \tag{3}$$

Then, use the residuals from this regression $\hat{\varepsilon}_2$ to run the final regression

$$y = \mathring{\beta}_0 + \mathring{\beta}_2 x_2 + \mathring{\beta}_3 (z + \hat{\varepsilon}_2) + \varepsilon_3, \tag{4}$$

where z is a constructed variable (possibly by statistical software) that has very weak (but non-zero) correlation with x_2.

All of the other (unknown and known) variables that explain y are included in $\hat{\varepsilon}_2$. Thus, this regression precludes spurious relationships, since all of the other (lurking) variables that explain y are accounted for in $\hat{\varepsilon}_2$.

Consequently, we can test for a strong, real causality as follows:

$$y(t) = \beta_0 + \beta_2 x_2 (t-1) + \beta_3 (z + \hat{\varepsilon}_2) + \varepsilon_3 (t), \tag{5}$$

where $x_2 (t-1)$ is the lag of x_2, and $\hat{\varepsilon}$ is defined as before. If $\beta_2 \neq 0$, then x_2 causes y.

Non-linear functional forms:

If the relationship between y and x_i is non-linear, we can use a Taylor expansion of the highest order. The model is linear in the parameters; therefore, our method can still be applied without any added complexity. This eliminates the specification error due to the functional form.

In sum, despite its simplicity, this method solves significant and longstanding problems in estimation and forecasting.

References

Alghalith M. (2018): The perfect regression and causality test: A solution to regression problems. Biometrical Letters 55: 45–48.

Alghalith M. (2019): A simple solution to the specification error. Biometrical Letters 56: 13–16. doi:10.2478/bile-2019-0002

Asteriou D., Hall S.G. (2011): Applied Econometrics. Palgrave MacMillan, London.

Golden R., Henley S., White H., Kashner T. (2016): Generalized information matrix tests for detecting model misspecification. Econometrics 4: 1–24.

Granger C.W.J. (1969): Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37: 424–438.

Maddala G.S., Lahiri K. (2009): Diagnostic checking, model selection, and specification testing. In: *Introduction to Econometrics* (4th ed.). Wiley, Chichester.

Phillips P.C.B. (1986): Understanding spurious regressions in econometrics. Journal of Econometrics 33: 311–340.

Shiffrin R.M. (2016): Drawing causal inference from big data. Proc Natl Acad Sci USA 113: 7308–7309.

Varian H.R. (2016): Causal inference in economics and marketing. Proc Natl Acad Sci USA 113: 7307–7315.