sciendo

DOI: 10.2478/bile-2019-0002

Biometrical Letters
Vol. 56 (2019), No. 1, 13-16

A simple solution to the specification error

Moawia Alghalith

Economics Dept., UWI, St Augustine, Trinidad and Tobago
e-mail: malghalith@gmail.com

Summary

We develop a simple method that completely eliminates the specification error and spurious relationships in regression. Furthermore, we introduce a stronger test of causality. We apply our method to oil prices.

Key words: spurious regression, causality, specification error, oil price

1. Introduction

Determining the true model has been a major obstacle in regression analysis. This is due to specification problems such as omitted variables, irrelevant variables, or the wrong functional form. It is virtually impossible for the researcher to know a priori all the relevant explanatory variables. The existing methods that deal with the specification error offer partial and limited solutions (see, for example, Golden et al., 2016, Maddala and Lahiri, 2009 and Wooldridge, 2013).

Needless to say, the misspecification of a model may cause spurious regressions due to the omission of lurking (confounding) variable(s) (see, for example, Phillips, 1986 and Asteriou and Hall, 2011). Moreover, this omission substantially weakens the existing causality tests, such as Granger's test and related tests (see Granger, 1969). These tests are weak and suffer well-known limitations (see, for example, Kleinberg, 2012 and Alghalith, 2018 for a discussion). Other causality tests also suffer limitations and are somewhat cumbersome (see, for example, Shiffrin, 2016 and Varian, 2016 for a discussion).

In this paper, we introduce a simple method that solves the problem of model misspecification. In doing so, we devise a simple method that completely eliminates the specification error. We also introduce a test of stronger causality (relative to Granger's causality).

2. The method

Choose any insignificant explanatory variable x_1 that is not correlated with the dependent variable y and run the 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

$$y = \beta_0 + \sum_i \beta_i x_i + \varepsilon_1, i \neq 1,$$

where clearly, $\varepsilon = \sum_i \beta_i x_i + \varepsilon_1$. Now, assume x_2 is the variable of interest. Then, we use the residuals of the regression in (1) $\hat{\varepsilon}$ to estimate the following regression:

$$y = \beta_2 + \beta_3 x_2 + \beta_4 \hat{\varepsilon} + \varepsilon_2. \tag{2}$$

All the other (unknown and known) variables that explain y are included in $\hat{\varepsilon}$. This is particularly helpful for models with lags, since the choice of the number of lags is a major difficulty in empirical research. According to our method, the researcher can choose one lag, while all the other relevant lags are automatically included in $\hat{\varepsilon}$. Clearly, the model specification in (2) precludes spurious relationships, since all the other (lurking) variables that explain y are accounted for in $\hat{\varepsilon}$.

Moreover, in general, the OLS estimates are unbiased, consistent and efficient, since the model is perfectly specified and generally there is no reason to believe otherwise. Consequently, we can test for a strong version of causality (relative to Granger's causality) as follows.

First, run the regression

$$y\,(_t) = \beta_5 + \beta_6 x_2\,(t-1) + \beta_7 \hat{\varepsilon} + \varepsilon_3\,(t)\,,$$

where $x_2\,(t-1)$ is the lag of x_2, and $\hat{\varepsilon}$ is defined as before. If $\beta_6 \neq 0$, then x_2 causes y.

3. Empirical example

As an example, we used a recent sample of monthly data (2000–2016) for U.S. oil production y, and the lag of the oil price $p(-1)$ as the explanatory variable of interest, while the futures price f is used as the (potentially)

insignificant explanatory variable. First, we estimated this regression (using OLS):

$$y = \beta_0 + \beta_1 f + \varepsilon. \tag{3}$$

Then, we used the residuals $\hat{\varepsilon}$ to estimate the following regression:

$$y = \beta_2 + \beta_3 p(-1) + \beta_4 \hat{\varepsilon} + \varepsilon_1. \tag{4}$$

The results appear in Table 1. As expected, from (4) , the adjusted R^2 is extremely close to one (0.9999) and the F-statistic = 667052.8 (a virtually perfect model). However, from (3) , adjusted $R^2 = -.001$ ($R^2 = .003$), and thus it badly failed the F-test (F-statistic = 0.74). All of the parameters in (4) are significant. Therefore, the oil price affects oil production.

Table 1. Empirical results (standard errors in parentheses)

β_0	182569.8	(7356.856)
β_1	92.26604	(106.6340)
β_2	182708.7	(89.78595)
β_3	90.25090	(1.300387)
β_4	0.998389	(0.000868)

4. Further refinements

We can also apply the method of Alghalith, 2018 to the residuals $\hat{\varepsilon}_1$ from (4) . In particular, we use (2) − (4) in Alghalith, 2018. This will result in virtually perfect results and solves other regression problems such as non-stationarity, heteroskedasticity, autocorrelation, and endogeneity. In addition, hypothesis testing becomes virtually redundant.

5. Non-linear functional forms

If the relationship between y and x_i is non-linear, we use the method in Alghalith, 2018 (repeated here for convenience); we obtain the following exact Taylor's expansion around a vector $\mathbf{c}$:

$$y = f(\mathbf{x}) = f(\mathbf{c}) + \sum\nolimits_i f_{x_i}(\mathbf{c})\,\bar{x}_i + R(\mathbf{x}, \mathbf{c}),$$

where R is the remainder and $\mathbf{x}$ is a vector of regressors. The remainder is given by

$$R(\mathbf{x}, \mathbf{c}) = \frac{1}{2} \sum f_{x_i x_j}(\dot{\mathbf{x}})\,\bar{x}_i^2 \bar{x}_j^2.$$

The remainder can be approximated as

$$R\left(\mathbf{x},\mathbf{c}\right)\approx\sum\beta_i\tilde{x}_i^2\tilde{x}_j^2,$$

where β_i is a parameter and the superscript $\sim$ denotes the deviation from the point of expansion. Therefore, we obtain the following regression model: $y=\beta_0+\sum\beta_i x_i+\sum\beta_i\tilde{x}_i^2\tilde{x}_j^2+\varepsilon_4$. The model is linear in the parameters; therefore, we can apply our method as before. This will eliminate the specification error due to the functional form (see Alghalith, 2018).

References

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

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.

Kleinberg S. (2012): Causality, Probability, and Time. Cambridge University Press, Cambridge.

Maddala G.S., Lahiri K. (2009): Diagnostic Checking, Model Selection, and Specification Testing. Introduction to Econometrics. 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. Proceedings of the National Academy of Sciences of the United States of America 113: 7308–7309.

Varian H.R. (2016): Causal inference in economics and marketing. Proceedings of the National Academy of Sciences of the United States of America 113: 7–7315.

Wooldridge J. (2013): Introductory Econometrics: A Modern Approach. South-Western Cengage Learning, Mason.