DE GRUYTER OPEN

DOI: 10.2478/bile-2018-0004

Biometrical Letters
Vol. 55 (2018), No. 1, 45-48

The perfect regression and causality test: A solution to regression problems

Moawia Alghalith

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

Summary

We introduce a method that eliminates the specification error and spurious relationships in regression. In addition, we introduce a test of strong causality. Furthermore, hypothesis testing (inference) becomes almost unneeded. Moreover, this method virtually resolves error problems such as heteroscedasticity, autocorrelation, non-stationarity and endogeneity.

Key words: spurious regression, strong causality, specification error

1. Introduction

Model specification has been a major obstacle in regression analysis. This is due to omitted variables, irrelevant variables, or the wrong functional form. It is virtually impossible for the researcher to know a priori all of the relevant explanatory variables. The existing methods that deal with specification error offer partial and limited solutions (see, for example, Asteriou and Hall, 2011).

Furthermore, the related problem of spurious regressions remains largely unresolved. Needless to say, the existing causality tests, such as Granger's test and the related literature (see Granger, 1969, 1980) are weak and suffer well-known limitations. Alternative methods of determining causality also suffer serious limitations and are cumbersome (see, for example, Shiffrin, 2016 and Varian, 2016 for discussion).

In this paper, we introduce a simple method that resolves the problems of specification error, spurious regressions, and weak causality. In doing so, we introduce a method that eliminates the specification error. We also develop a test of strong causality.

2. The method

Assume the true model is

$$y = \beta_0 + \beta_1 x_1 + \sum\nolimits_i \beta_i x_i + \varepsilon,$$

where β is the parameter to be estimated, and ε is the error. We adopt the following two-step estimation procedure. First, we estimate the following regression (assuming that x_1 is the explanatory variable of interest):

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

where u is the error; clearly, $u = \sum \beta_i x_i + \varepsilon$. We fit the line of regression (1) and then we use the residuals of this regression $\hat{u}$ to accurately estimate the true parameter. We note that all of the (unknown and known) explanatory variables that explain y are included in $\hat{u}$. The residuals are given by

$$\begin{aligned} \hat{u} &= y - \hat{\beta}_0 - \hat{\beta}_1 x_1 = \beta_0 + \beta_1 x_1 + \sum\nolimits_i \beta_i x_i + \varepsilon - \hat{\beta}_0 - \hat{\beta}_1 x_1 \\ &= \beta_0 - \hat{\beta}_0 + \left(\beta_1 - \hat{\beta}_1\right) x_1 + u, \end{aligned}$$

where $\hat{\beta}$ is the estimated parameter. Multiplying both sides by x_1 yields

$$\hat{u} x_1 = (\beta_0 - \hat{\beta}_0) x_1 + \left(\beta_1 - \hat{\beta}_1\right) x_1^2 + \varepsilon_1,$$

where $\varepsilon_1 = u x_1$. We save the residuals from the regression $\hat{\varepsilon}_1$ and use them as a new explanatory variable in the original regression in (1), as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 \frac{\hat{\varepsilon}_1}{x_1} + \varepsilon_2, \tag{2}$$

where $\frac{\hat{\varepsilon}_1}{x_1}$ is an estimate of $u = \sum \beta_i x_i + \varepsilon$, since $u = \frac{\varepsilon_1}{x_1}$. The fitted line of this regression is given by

$$y = \beta_0^* + \beta_1^* x_1 + \beta_2^* \frac{\hat{\varepsilon}_1}{x_1} + \varepsilon_2^*. \tag{3}$$

Subtracting (3) from (2) yields

$$Z = \beta_0 - \beta_0^* + (\beta_1 - \beta_1^*) x_1 + \beta_2 \frac{\hat{\varepsilon}_1}{x_1} + \varepsilon_2 - \varepsilon_2^*,$$

where $Z \equiv \beta_2^* \frac{\hat{\varepsilon}_1}{x_1}$. Then we estimate this regression

$$Z = \beta_0 - \beta_0^* + (\beta_1 - \beta_1^*) x_1 + \beta_2 \frac{\hat{\varepsilon}_1}{x_1} + \varepsilon_3,$$

where the error $\varepsilon_3 = \varepsilon_2 - \varepsilon_2^*$. We note that the residuals from this regression $\hat{\varepsilon}_3$ are an estimate of $\varepsilon_2 - \varepsilon_2^*$ (the error of the error). We use $\hat{\varepsilon}_3$ to re-estimate (2) as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_3 v + \varepsilon_4, \tag{4}$$

where $v \equiv \beta_2^* \frac{\hat{\varepsilon}_1}{x_1} + \hat{\varepsilon}_3$.

We note that all of the explanatory variables that explain y are included in v. Therefore the model is perfectly specified. This virtually resolves the (choice of variable) misspecification problem. This is particularly helpful for models with lags, since the choice of number of lags is a major difficulty in empirical studies. According to our method, the researcher can choose one lag, while all other relevant lags are automatically included in $\hat{u}$. We also note that hypothesis testing (and sampling) becomes almost uneeded, since the residuals are virtually zero, regardless of the sample choice or size, and thus the parameters are virtually equal to the true parameters. Furthermore, this method virtually resolves error problems such as heteroscedasticity, autocorrelation, non-stationarity and endogeneity. We also note that, if needed, the procedure (2) – (4) can be repeated until the standard error of the regression is zero. A computer program can easily and quickly achieve this.

3. Large causality test

Clearly, the model specification in (4) precludes spurious relationships, since all of the other variables that explain y are accounted for in $\hat{u}$. Consequently, we can test for a strong version of causality as follows:

First, run these regressions

$$y_t = \beta_4 + \beta_5 x_{t-1} + \varepsilon_4,$$

$$x_t = \beta_6 + \beta_7 y_{t-1} + \varepsilon_5,$$

where the subscript $t-1$ denotes the first lag. Follow the procedure in the previous section to obtain

$$y_t = \beta_4 + \beta_5 x_{t-1} + \beta_8 v_1 + \varepsilon_6,$$

$$x_t = \beta_6 + \beta_7 y_{t-1} + \beta_9 v_2 + \varepsilon_{7,}$$

then estimate these regressions. If $\beta_5 \neq 0$ and $\beta_7 = 0$, we conclude that x causes y.

4. A non-linear functional form

If the relationship between y and x_i is non-linear, we obtain the following exact Taylor's expansion around a vector of constants $\mathbf{c}$:

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

where R is the remainder and $\mathbf{x}$ is a vector of regressors, f_{x_i} is the first partial derivative of f with respect to x_i, and the bar superscript denotes the deviation from the point of expansion ($\bar{x}_i \equiv x_i - c_i$). The remainder is explicitly given by

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

where $\dot{x}_i$ is a number between x_i and c_i, and $\dot{\mathbf{x}}$ is a vector of these numbers. The remainder can be approximated as

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

where β_i is a parameter. Thus, we obtain the following regression model

$$y = \beta_0 + \sum \beta_i x_i + \sum \beta_i \bar{x}_i^2 \bar{x}_j^2 + \varepsilon_4.$$

Now 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.

References

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

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

Granger C.W.J. (1980): Testing for causality: a personal viewpoint. Journal of Economic Dynamics and Control 2: 329–352.

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: 7310–7315.