A new insight of specification error in regression: Excluding education variable from a model

by Andryan Setyadharma

Submission date: 20-Sep-2023 08:48AM (UTC+0700)
Submission ID: 2171159507
File name: 36._A_new_insight_of_specification_error.pdf (800.77K)
Word count: 3288
Character count: 17943

Journal of Physics: Conference Series

A new insight of specification error in regression: Excluding education variable from a model

A new insight of specification error in regression: Excluding education variable from a model

A Setyadharma[1], P A Bowo[1], and D A Suseno[1]
[1] Development Economics Department, Universitas Negeri Semarang, Semarang City, Central Java Province, Indonesia

Email: andryan@mail.unnes.ac.id

Abstract. The aim of this study is to show specification error in a model that exclude a relevant independent variable. Literature suggests that human capital features, such as education, must be considered in studies of social problems, such as environmental degradation. Therefore, the absence of education variable in a multiple regression equation model that explain the determinants of environmental degradation may result in model misspecification. Firstly, this study constructs a model with four variables, i.e.: unemployment level (ULEVEL), poor areas (POOR_AREAS), income inequality (INEQ) and access to electricity (ELEC), that may affect environmental quality in Indonesia. Secondly, this study examines the second model when education variable is included in the model. This study uses panel data regression method from 33 provinces in Indonesia during 2012 to 2018. The results of the first model show that all four control variables are statistically significant. However, ELEC's coefficient has unexpected sign. When education variable is added to the second model, ELEC's coefficient has expected sign. In conclusion, education variable is theoretically and empirically important variable to explain the changes in the quality of environment in Indonesia and education variable should be used in the model to avoid specification error in a model.

1. Introduction

Econometrics is defined as the use of mathematics and statistical methods in the study of economic data as proxied in specific economic model[1]. Furthermore, [1]suggest that researchers in classical econometrics usually work with models such as regression models where they use data dan estimate them with the application of the correct technique and provide the coefficients of the model.[2] proposes two main aims of a regression model are to explain the variability of an explanatory variable; and to forecast future values of the explanatory variable. There are three steps of regression analysis: the model specification, the estimation of the coefficients of this model, and the interpretation of these coefficients . [3]argues that model specification is one of the most vital issues in regression analysis but get less attention from the researchers and he also argues that the correct model specification is the most important step because the estimation result and correct interpretation of the coefficients of the model are subject to the correct specification of the model.

Model specification refers to the decision of inclusion or exclusion of independent variables from a multiple regression equation. The inclusion of independent variables in a multiple regression model is generally based on theoretical concerns rather than empirical or methodological aspects. In fact, a multiple regression model reflects a theoretical statement about the causal relationship between one or more explanatory variables and a dependent variable[3]. Then, researchers deduct the hypothesis based on theories that is relevant to the multiple regression model [4]. However, researchers sometime decide

to remove some of the independent variables that are stipulated in theories due to the difficulties to quantify the variables or to get proper observations on the variables. The omission of the important variables in the theories may result in specification error.

Problems may occur when a model is wrongly specified. [1] argue that exclusion of relevant variables creates biased and inconsistent parameters due to high variances and standard errors. [1] also argue that inclusion of unrelated variables causes larger estimated variance than it should be. Model misspecification may cause estimation and interpretation problems. In consequence, estimation of the model may generate outputs that are incorrect or misleading. For example, [5] suggests that exclusion of relevant variables causes wrong sign of regression coefficients. Wrong sign of regression coefficients makes wrong interpretation of regression results.

The aim of this study is to show specification error in a model that exclude a relevant explanatory variable, or the Type B error. This study demonstrates the problems association with Type B error using a causal relationship between education variable and quality of environment variable in case of Indonesia. It is argued that human capital features, such as education, must be considered in studies of social problems [6]. Environmental problems are not only limited to pollution problem, but it is also a social problem because it is a problem that endanger the current patterns of social organization and social thought [7]. Therefore, the absence of education variable in a regression equation model that explain the determinants of environmental degradation may result in model misspecification due to mis-specify the model by omitting an explanatory variable that is theoretically relevant.

This paper adds to body of knowledge on the importance of inclusion of education variable to explain environmental degradation in the case of Indonesia. This paper examines an example of the problem related with the omission of a theoretically relevant explanatory variable from a multiple regression equation because [3] argue that it is more serious than the problems linked with the addition of a theoretically irrelevant explanatory variable. The paper has five sections. Section 2 briefly examines the conceptual relationship between education and quality of environment. Section 3 explains the research method and discussion the results is presented in Section 4 and Section 5 is the conclusion.

2. The Relationship between Education and Quality of Environment

As mentioned earlier, the decision to include independent variables in a multiple regression model should be based on theories rather than empirical or methodological aspects. In this section, we discuss the theoretical background about the relationship between education and environmental degradation. It becomes an evident that human is responsible to the damage of the environment in the name of development. Recent views state that education has been named as an essential factor to the protection of the environment.

[8] argues that poor people with low levels of education usually do not understand the consequences of their activities that cause into the worsening of environmental quality. Impoverished people with low education levels mainly focus on their basic needs for survival first instead of the preservation of the environment so they have been regularly blamed of the low quality of the environment. A study by [9] states that there are three motives behind this relationship, i.e.: (a) the poor people greatly count on natural resources for their daily food source; (b) poor people only concern about present benefits rather than future benefits, and (c) poor people have less accesses to resources. Increased access to education is not only to contribute on reducing poverty but also improve their awareness of the protection of the environment.

[10] propose there are three explanations for positive relation between education and environmental sustainability. First, individuals with high education level are more reactive of environmental crises and they have greater preferences for better quality of environment and always behave to protect the environment. Second, people with higher education intend and able to utilize the available ways to display their environmental choices, set up advocacy groups, participate in them, and share their ideas about public environmental policies. Third, people with high school levels can force governments and other stakeholders to more open minded to public demands for the improvement of environmental quality. However, undeniably, education can also turn to be the opposite, where a higher educational

attainment might worsen the environmental quality through income channel. Educated individuals usually earn more income and they may demand for more goods that do not eco-friendly.

In conclusion, literature has shown that education has significant effect on the people's environmentally lifestyles and behavior. A higher level of people's education is likely also enhancing their request for a higher level of environmental protection. Therefore, the inclusion of education variable in a regression equation model that explain the determinants of the deterioration of the environment is a mandatory condition. In next section, we show the consequence of eliminating a relevant explanatory variable on the partial regression parameter of the other explanatory variable in a regression equation.

3. Research Methods

This study applies panel data technique. The data are annual and secondary data recorded from 2012 to 2018 from 33 out of 34 Provinces in Indonesia. One province, i.e. Kalimantan Utara Province, is excluded due to inadequate dataset. Kalimantan Utara Province is a new province, and the environment quality data of this province was not available until 2017. This study sets two equation models to examine the consequence of eliminating education variable as one of the determinants of environmental degradation. Firstly, a model with four explanatory variables (i.e. unemployment Level (ULEVEL), percentage of poor areas (POOR_AREAS), income inequality (INEQ) and percentage of zones with electricity (ELEC)), is constructed to explain the changes in environmental quality (ENVI) in Indonesia. The first model is set as follow:

$$L(ENVI)it = \alpha 0 + \alpha 1 ULEVELit + \alpha 2 POOR\_AREASit + \alpha 3 L(INEQ)it + \alpha 4 ELECit + eit \quad (1)$$

Where, L represents the function of logarithm, e stands for the disturbance which is assumed has a normal distribution, i denotes the cross-sections on provincial level and t represents the time, $\alpha 0$ is a constant, $\alpha 1$, $\alpha 2$, $\alpha 3$, $\alpha 4$ are the coefficients of the First Model. The dataset of ENVI are minimum of 0 (zero) up to maximum of 100, where 0 is the worst value and 100 is the best value, considers values closer to 100 is better quality. It has opposite way than the usual environmental quality measurement. Secondly, the first model is added with education (EDUC) variable using Expected Years of Schooling per province to represent the educational attainment level. The second model is constructed as follow:

$$L(ENVI)it = \beta 0 + \beta 1 ULEVELit + \beta 2 POOR\_AREASit + \beta 3 L(INEQ)it + \beta 4 ELECit + \beta 5 L(EDUC)it + uit \quad (2)$$

Where, u represents for the disturbances of the second model, $\beta 0$ is a constant, $\beta 1$, $\beta 2$, $\beta 3$, $\beta 4$ and $\beta 5$ symbolize the parameters of the second model. Before analyzing the panel data outputs, the initial procedures of panel data is the choice of the appropriate model. There are three types of data panel, i.e., Common Effect Model (CEM), Fixed Effect Model (FEM) and Random Effect Model (REM) and the best model is selected base on three tests: Chow test, Lagrange Multiplier test and Hausman test.

4. Results and Discussion

The assessment tools of the best model indicate that Fixed Effect Model is the best for First Model. Both Chow Test and Hausman Test specify that that the better model is Fixed Effect Model.

Tests	Statistic Values	*p*-values
Chow *Test*	34.77	0.0000 (Rejects CEM, FEM is selected))
LM *Test*	419.46	0.0000 (Reject CEM, REM is selected))
Hausman *Test*	21.27	0.0003 (Reject REM, FEM is selected)
Decision	**FEM is selected as the appropriate model**	

Table 1. The Selection of Finest Model of the First Model

Table 2 shows the result of the fixed effect model of the First Model. To reduce heteroscedasticity problems in the model, the First Model is estimated with White Cross-section standard errors & covariance procedure. The output in table 2 indicates that ULEVEL negatively associated on ENVI while POOR_AREAS and INEQ also have a negative influence on ENVI. Only ELEC has positive effect on ENVI in this model. All explanatory variables are significantly affecting the explained variable. The signs of regression coefficients of ULEVEL, POOR_AREAS and INEQ are as expected, while the sign of regression coefficient of ELEC is unexpected.

Independent Variables	**Coefficients**
ULEVEL	-0.013 [-3.314]***
POOR_AREAS	-0.003 [-4.363]***
LOG (INEQ)	-0.200 [-1.837]*
ELEC	0.004 [2.980]***
Constant	3.732 [37.297]***
Adjusted R^2	0.903

Note: **Explained Variable: Log of ENVI**
values in parentheses are t-statistics values. *** significance at $p \leq 0.01$; ** significance at $p \leq 0.05$; * significance at $p \leq 0.10$

Table 2. Result of the FEM of the First Model

It seems that the regression coefficient of ELEC has a wrong sign. The wrong sign in this case means that the regression coefficient of ELEC has unexpected sign. This result indicates that if the ELEC increases then the ENVI also increases, holding other variables constant. In other words, if the percentage of zones with electricity increases then the Environmental Index also increases. This results seem contradicts because [11] indicates that the major contributor for Green House Gas emissions in Indonesia are CO_2 emissions from energy, in which the share of electricity, heat and other to total CO2 emissions in 2017 is 36%. The share is the highest among the others[12]. It means that higher percentage of zones with electricity in Indonesia will deteriorate the environmental quality. As suggested by [5], the wrong sign of the regression coefficient of ELEC indicates that there is specification error in First Model due to the omission of an explanatory variable that is theoretically relevant.

The next step is the evaluation of the Second model.[13] The second model is the First Model with addition of Education variable. The best model choice of panel data of Second Model indicates that the best model is also FEM, as presented in Table 3.

Tests	**Statistic Values**	***p*-values**
Chow *Test*	49.03	0.0000 (Rejects CEM, FEM is selected))
LM *Test*	434.38	0.0000 (Reject CEM, REM is selected))
Hausman *Test*	12.06	0.0339 (Reject REM, FEM is selected)
Decision	**FEM is selected as the appropriate model**	

Table 3. The Selection of Finest Model of the Second Model

As shown in Table 4, the Second Model is estimated with White Cross-section standard errors & covariance procedure to reduce the presence of heteroscedasticity issues. There is no different of the signs of regression coefficients of ULEVEL, POOR_AREAS and INEQ in Second Model in comparison to First Model. However, the sign of regression coefficient of ELEC has switched in Second Model.

Now it is a negative sign. The inclusion of EDUC variable has changed the sign of regression coefficient of ELEC variable. It is concluded that the inclusion of an explanatory variable that is theoretically relevant (in this case, EDUC variable) has avoided a specification error problem in the model. EDUC itself has expected sign, i.e. positive coefficient. EDUC is statistically significance t-statistics which it clearly confirms the positive effect of education on the quality of environment in Indonesia.

Independent Variables	**Coefficients**
ULEVEL	-0.007 [-1.923]*
POOR_AREAS	-0.003 [-3.741]***
LOG (INEQ)	-0.090 [-2.128]**
ELEC	-0.003 [-2.362]**
LOG (EDUC)	0.885 [5.985]***
Constant	2.250 [6.229]***
Adjusted R^2	0.918

Note: **Explained Variable: Log of ENVI**
values in parentheses are t-statistics values. *** significance at $p \leq 0.01$; ** significance at $p \leq 0.05$; * significance at $p \leq 0.10$

Table 4. Result of the FEM of the Second Model

5. Conclusion

The aim of this study is to show specification error problem when a relevant independent variable is excluded from a model. Based on the findings, it is confirmed that the specification error occurs by omitting an explanatory variable that is theoretically relevant. In the beginning, we employ three tests to choose the better models (CEM, FEM or REM) for First Model and Second Model. The results indicate that FEM is the best model for both models. The regression output of First Model shows that all four control variables are statistically significant. However, ELEC's coefficient has unexpected sign and when education variable is added to the model, ELEC's coefficient has converted into expected sign. In addition, Second Model also confirms the significant impact of educational level on the quality of environment. It is inferred that higher educational level is related with better quality of environment. However, although education is taking a big part of the solution of the better quality of environment, it is important to note that environmental issues are never simply can be solved by better education only. It also requires support from other factors such as economic, political, and social aspects.

References

[1] Andrikopoulos A A and Gkountanis D C 2011 Issues and Models in Applied Econometrics: a Partial Survey *South-Eastern Eur. J. Econ.* **9** 107–65
[2] Deegan Jr. J 1976 The Consequences Of Model Misspecification In Regression Analysis *Multivariate Behav. Res.* **11** 237–48
[3] Anon Model specification in regression analysis *Understanding Regression Analysis* (Boston, MA: Springer US) pp 166–70
[4] Banno K, Ramsey C, Walld R and Kryger M H 2009 Expenditure on health care in obese women with and without sleep apnea *Sleep* **32** 247–52
[5] Mullet G M 1976 Why Regression Coefficients Have the Wrong Sign *J. Qual. Technol.* **8** 121–6
[6] Jorgenson A K 2003 Consumption and environmental degradation: A cross-national analysis of the ecological footprint *Soc. Probl.* **50** 374–94
[7] Bell M M 1998 An Invitation to Environmental *Sociol. Thousand Oaks, CA Pine Forge Press*

[8] Sobhee S K 2004 Economic development, income inequality and environmental degradation of fisheries resources in Mauritius *Environ. Manage.* **34** 150–7
[9] Zhen N, Fu B, Lu Y and Wang S 2014 Poverty reduction, environmental protection and ecosystem services: A prospective theory for sustainable development *Chinese Geogr. Sci.* **24** 83–92
[10] Farzin Y H and Bond C A 2006 Democracy and environmental quality *J. Dev. Econ.* **81** 213–35
[11] Mensah G A, Cooper R S, Siega-Riz A M, Cooper L A, Smith J D, Brown C H, Westfall J M, Ofili E O, Price L N and Arteaga S 2018 Reducing cardiovascular disparities through community-engaged implementation research: a National Heart, Lung, and Blood Institute workshop report *Circ. Res.* **122** 213–30
[12] Budiono E 2019 Pengaruh kedisiplinan , perhatian orang tua dan jumlah saudara terhadap prestasi belajar matematika Influence of discipline , parent ' s attention and number of brothers on mathematics learning achievement **1** 16–22
[13] Muzayyanah A and Wutsqa D U 2019 Annals of Mathematical Modeling , 1 (2), 2019 , 47-63 Effectiveness of problem posing and investigation in terms of problem solving abilities , motivation and achievement in mathematics **1** 47–63

A new insight of specification error in regression: Excluding education variable from a model

ORIGINALITY REPORT

7%	5%	5%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ueaeprints.uea.ac.uk Internet Source	1%
2	George A. Mensah, Richard S. Cooper, Anna Maria Siega-Riz, Lisa A. Cooper et al. "Reducing Cardiovascular Disparities Through Community-Engaged Implementation Research", Circulation Research, 2018 Publication	1%
3	ejournal.unsri.ac.id Internet Source	1%
4	Michinori Honma, Toshiaki Nose, Susumu Sato. "Improvement of Aberration Properties of Liquid Crystal Microlenses using the Stacked Electrode Structure", Japanese Journal of Applied Physics, 2001 Publication	1%
5	Gallego-Alvarez, Isabel, Mª Vicente-Galindo, Mª Galindo-Villardón, and Miguel Rodríguez-Rosa. "Environmental Performance in Countries Worldwide: Determinant Factors	1%

and Multivariate Analysis", Sustainability, 2014.
Publication

6	technodocbox.com Internet Source	1 %
7	www.proceedings.com Internet Source	1 %
8	Jiwhan Noh, Jae-Hoon Lee, Sang-Yup Lee, Suckjoo Na. "Fabrication of Random Microspikes on Mold Metal by Ultrashort Laser Ablation for Hydrophilic Surface", Japanese Journal of Applied Physics, 2010 Publication	1 %
9	jim.ar-raniry.ac.id Internet Source	<1 %
10	iopscience.iop.org Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		

A new insight of specification error in regression: Excluding education variable from a model

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7