Bivariate regression model and its applications

by Budi Pratikno

The bivariate regression model and its application

The bivariate regression model and its application

B Pratikno[*], **L Sulistia and Saniyah**

Department of Mathematics, Faculty of Mathematics and Natural Science, Jenderal Soedirman University, Purwokerto, Indonesia

[*]Corresponding author: bpratikno24@gmail.com

Abstract. The paper studied a bivariate regression model (BRM) and its application. The maximum power and minimum size are used to choose the eligible tests using non-sample prior information (NSPI). In the simulation study on real data, we used Wilk's lamda to determine the best model of the BRM. The result showed that the power of the *pre-test-test* (*PTT*) on the NSPI is a significant choice of the tests among *unrestricted test* (*UT*) and *restricted test* (*RT*), and the best model of the BRM is $Y_{(1)} = -894 + 46X$ and $Y_{(2)} = 78 + 0.2X$ with significant Wilk's lamda 0.88 < 0.90 (Wilk's table).

1. Introduction

The paper studied a bivariate regression model (BRM) and its application with and without non-sample prior information (NSPI). In the context of the testing hypothesis with NSPI, the power and size of the tests unrestricted test (UT), restricted test (RT) and pre-test-test (PTT) are used. Many authors have already studied about UT, RT, PTT in testing intercept using non-sample prior information (NSPI) such as Pratikno (2012), Khan and Pratikno (2013), Yunus and Khan (2011), Saleh and Sen (1983) and Khan, et al. (2016). Some previous research related to preliminary testing and NSPI are also found in Han and Bancroft (1968), Judge and Bock (1978), and Saleh (2006), Tamura (1965), and Saleh and Sen (1978, 1982). Other authors such as Khan (2005, 2008), Khan and Saleh (1997, 2005, 2008), Khan and Hoque (2003), Saleh (2006) and Yunus (2010) contributed to this research topic in the context of estimation area, and Tamura (1965), Saleh and Sen (1978, 1982), and Yunus and Khan (2011) discussed testing hypothesis with NSPI on the non parametric model. All the authors used *R-code* in *R-package* especially mvtnorm and cubature function, respectively. To choose the best test, we choose a maximum power and minimum size of the tests (UT, RT, PTT) on multivariate simple regression model (MSRM), especially on a bivariate simple regression model (BSRM). Note that the definition of the power is probability reject H_0 under H_a in testing $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$, and the power should close to 1 under H_a and tend to be 0 under H_0 (Casella and Berger, 2002: 383).

In the simulation on real data, the BRM (or BSRM with single predictor) requires the correlation between responses. Here, the *Wilk's Lamda* and mean deviation error (MDE) are used to check the best model of the BRM and or BSRM. Following Rencher (2002), the general model of the BSRM is given by

IOP Conf. Series: Journal of Physics: Conf. Series **983** (2018) 012109 doi:10.1088/1742-6596/983/1/012109

$$\mathbf{Y}_{n\times 2} = \mathbf{X}_{n\times(q+1)}\mathbf{B}_{(q+1)\times 2} + \boldsymbol{\varepsilon}_{n\times 2} \tag{1}$$

With $\boldsymbol{Y}$ is a bivariate response, $\boldsymbol{X}$ is predictors and $\boldsymbol{\varepsilon}$ is error term. To test intercept with NSPI and create the best model, the power and size of the tests and MDE are used, respectively. A Simulation study is done on climate data, namely temperature (X), rainfall (Y_1), and humidity (Y_2). The research is done in some steps: (1) we firstly review some previous research related to the power of the tests (UT, RT, PTT) on testing intercept with NSPI on regression model, (2) we do graphical analysis of the power of the tests in a simulation study with generating data from *R,* and (3) we finally create and analyse the best model of the BSRM for climate data in Cilacap, January 2009 to February 2014).
The research presented the introduction in Section 1. Testing with NSPI on the BSRM is given in Section 2. A simulation study to determine the best model of the BSRM on real data is then obtained in Section 3, and Section 4 described conclusion.

2. Methods

2.1. Testing with NSPI on the BSRM

Following Casella and Berger (2002), a power is probability reject H_0 under H_a in testing $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$. Similarly, the power of tests of the UT, RT, and PTT are given as probability reject H_0 under H_a in testing intercept using NSPI on MSRM and BSRM model. Therefore, we then refer to Pratikno (2012) for presenting power and size of the tests of the UT, RT, and PTT on the MSRM model. The results showed that the formula of the power and size of the tests for three tests UT, RT and PTT on BSRM (bivariate response, *p=2.*) model follows the MSRM model on Pratikno (2012). Furthermore, we use the bivariate noncentral *F* distribution to compute power and size of the PTT. Detail proposed tests, power and size of the tests and bivariate noncentral *F* are found in Pratikno (2012). In this simulation study, we generated random data using *R package*. The predictor is generated from the uniform distribution, $U \sim (0,1)$ and the errors are generated from the normal distribution, $Z \sim (0,1)$ In each case $n = 30$, random variates were generated. For the computation of the power and size of thee of the tests (UT, RT, and PTT) on the BSRM, we set $\alpha = 0.05$. We then present the power and size of the tests in Figure 1.

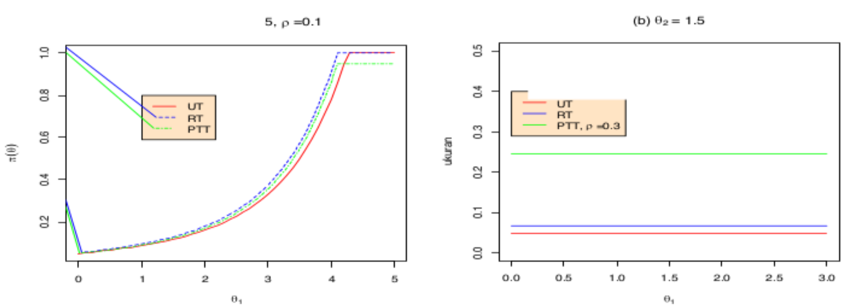

Figure 1. (i) Power UT, RT, and PTT with $\theta_2 > 0$ and $\rho = 0.1$; (ii) S*ize* UT, RT, and PTT with $\theta_2 = 1.5$

From Figure 1 (i), we see that the power of the PTT lies between the power of the UT and RT for $\rho = 0.1$. They start from the same value, that is 0.05, and they then increase as θ_2 increases. We also see that the size of the PTT maximum and UT minimum for $\theta_2 > 0$ (Figure 1,(ii)). It is clear that PTT is not a significant choice, but we still recommend PTT as a choice, as well as RT. This condition does not follow the previous research and theory. Furthermore, we recommend changing the distribution of the generated data for both predictor and error.

2.2. A Simulation Study on Real Data

Following Pratikno (2012), the MSRM is a model of multiple responses with a single predictor, and for $p = 2$ the MSRM become the BSRM. Furthermore, the general model of the MSRM of n observation of random variables with single predictor and vector response, $(X_i, \boldsymbol{Y}_i)$, $i=1,2,\ldots,n$, is given by

$$\boldsymbol{Y}_i = \beta_0 + \beta_1 X_i + e_i \tag{2}$$

With $\boldsymbol{Y}_i = (Y_{i1},\cdots,Y_{ip})^t$ is p dimension vector response, $e_i = (e_{i1},\cdots,e_{ip})$ is p dimension vector *errors* that follow normal distribution with *mean* **0** and covariance $\sum$, denoted by $e_i \propto N_p(0,\sum)$, X_i is a predictor, and $\boldsymbol{\beta_0} = (\beta_{01},\cdots,\beta_{0p})^t$ and $\boldsymbol{\beta_1} = (\beta_{11},\cdots,\beta_{1p})^t$ are vector intercept and slope parameters, respectively. The properties and assumption of the equation (2) are: (1) $E(\mathbf{Y}_{(i)}) = \mathbf{X}\boldsymbol{\beta}_{(i)}$ or $E(\boldsymbol{\varepsilon}_{(i)}) = \mathbf{0}$, (2) $Cov(\mathbf{y}_i) = \boldsymbol{\Sigma},$ for all $i = 1,2,\ldots,n$, (3) and $Cov(\mathbf{y}_i, \mathbf{y}_j) = \mathbf{0},$ all $i \neq j$ (Rencher, 2002:338-339).

To create the eligible model on the BSRM, we must check the correlation between responses. Let, $\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}, \ldots, \mathbf{Y}_{(p)}$ are response variables, then they are independent if

$$\chi_h^2 = -\left\{n - 1 - \frac{2p+5}{6}\right\} \ln|\mathbf{R}| , \tag{3}$$

$\chi_h^2 > \chi^2_{\alpha,\frac{1}{2}p(p-1)}$, . It means that there is a correlation between responses. Following Morrison (2005), this test is called *Bartlett's Test* with $|\mathbf{R}|$ is determinant of the correlation matrix responses, p is some response and $\frac{1}{2}p(p-1)$ is degrees of freedom. The *Wilk's lamda* and mean deviation error (MDE) are then used to detect the best model. Let, we rewrite again of the general model of the BSRM in the equation (1), namely $\mathbf{Y} = \mathbf{XB} + \boldsymbol{\varepsilon}$,

where

$$\mathbf{Y}_{(nx2)} = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \\ \vdots & \vdots \\ Y_{q1} & Y_{q2} \end{bmatrix}; \quad \mathbf{X}_{nx(q+1)} = \begin{bmatrix} 1 & X_{11} \\ 1 & X_{21} \\ \vdots & \vdots \\ 1 & X_{n1} \end{bmatrix}; \quad \mathbf{B}_{(q+1)x2} = \begin{bmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \\ \vdots & \vdots \\ \beta_{q1} & \beta_{q2} \end{bmatrix}; \quad \boldsymbol{\varepsilon}_{(nx2)} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} \\ \varepsilon_{21} & \varepsilon_{22} \\ \vdots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} \end{bmatrix}$$

with $E(\boldsymbol{\varepsilon}) = \mathbf{O}$ and $\text{cov}(\text{y}_i, \text{y}_j) = \mathbf{O},$ where O is null matrices, , for $i, k = 1, 2, ..., p$.

3. Result and discussion

To set the model of the BSRM above, we try and analyze two types of data as follows. Firstly, we use data of water discharge (X), sedimentation rate (Y_1) and *outflow* (Y_2). In this case, the independence of the response variable $(\mathbf{Y_1}, \mathbf{Y_2})$ is tested using *Bartlett test* in the equation (3). The result showed that they are not significant correlation. Therefore, we then used another data (climate data) to simulate the best model of the BSRM. Here, the responses have a significant correlation From this data, we obtain the value of the statistics test, $\chi_h^2 = 25.8$ (Saniyah and Pratikno, 2014) and it is greater than $\chi_{0,05;1}^2 = 3.8$. So we reject H_0, it means that the response has a significant correlation. Now, we are allowed to set the model of the BSRM from this data. Following Saniyah and Pratikno (2014), the linear model of the BSRM is $\mathbf{Y}_{n\times 2} = \mathbf{X}_{n\times(q+1)}\mathbf{B}_{(q+1)\times 2} + \boldsymbol{\varepsilon}_{n\times 2}$, with parameters estimate $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. We then obtained parameters estimate $\hat{\mathbf{B}} = \begin{bmatrix} -894 & 78 \\ 46 & 0.2 \end{bmatrix}$. Moreover, the statistics test Wilk's lamda (Λ_h) is then used to testing the parameters estimate. Here, we get $\Lambda_h = 0.88$ with $\Lambda_{table} = \Lambda_{0,05;2;1;60} = 0.90$ (0.88 < 0.90), and $MD_E = 1.2$. Finally, we conclude that the best model of the BSRM are $Y_{(1)} = -894 + 46X$ and $Y_{(2)} = 78 + 0.2X$.

4. Conclusion

The paper studied bivariate regression model (BRM) and its application with and without non-sample prior information (NSPI). The maximum power and minimum size (using NSPI) and (2) Wilk's lamda and mean deviation error (MDE) are used. The result showed that the power of the *pre-test-test* (PTT) could be a choice of the tests among *unrestricted test* (UT) and *restricted test* (RT), and the best model of the BSRM of climate data are $Y_{(1)} = -894 + 46X$ and $Y_{(2)} = 78 + 0.2X$, with *significant* MDE 1.2 and Wilk's lamda 0.88 < 0.90 (Wilk's table).

References

[1] Saleh A K Md E 2006 *Theory of preliminary test and Stein-type estimation with applications.* (New Jersey: John Wiley and Sons, Inc.)
[2] Rencher A C 2002 *Methods of Multivariate Analysis*. (2nd ed. New York: John Wiley and Sons Inc.)
[3] Saleh A K Md E and Sen P K 1978 *Ann. Stat.* **6** 154
[4] Saleh A K Md E and Sen P K 1982 *Commun. Stat.-Theory Methods* **12** 1855
[5] Saleh A K Md E and Sen P K 1983 *Commun. Stat.-Theory Methods* **11** 639
[6] Pratikno B 2012 *Tests of hypothesis for linear regression models with non sample prior information* Dissertation University of Southern Queensland, Australia
[7] Casella G and Berger R L 2002 *Statistical inference* (2nd Ed. Massachussets: Thomson Learning Inc.)
[8] Han C P and Bancroft T A 1968 *J. Am. Stat. Assoc.* **63** 1333
[9] Judge G G and Bock M E 1978 *The Statistical implications of pre-test and stein-rule estimators in econoetrics* (New York: North-Holland)
[10] Yunus R M and Khan S 2011 *Appl. Math. Comput.* **217** 6237
[11] Yunus R M 2010 *Increasing power of M-test through pre-testing* Unpublished PhD Thesis *University* of Southern Queensland Australia
[12] Khan S and Pratikno B 2003 *Stat. Pap.* **54** 605
[13] Khan S 2005 *J. Stat. Res.* **39** 79
[14] Khan S 2008 *Commun. Stat.-Theory Methods* **37** 247
[15] Khan S and Saleh A K Md E 1997 Biom. J. **39** 1
[16] Khan S and Hoque Z 2003 *J. Stat. Res.* **37** 43
[17] Khan S and Saleh A K Md E 2005 *Stat. Pap.* **46** 379
[18] Khan S and Saleh A K Md E 2008 *Commun. Stat.-Theory Methods* **37** 2564
[19] Saniyah and Pratikno B 2014 *Jurnal Ilmiah Matematika dan Pendidikan Matematika* **6** 45
[20] Tamura R 1965 *Bull. Math. Stat.* **11** 38

Bivariate regression model and its applications

ORIGINALITY REPORT

9%	%	%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universitas Diponegoro Student Paper	9%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		