Available online at www.sciencedirect.com

ScienceDirect

Procedia Economics and Finance 32 (2015) 68 – 76

Procedia Economics and Finance

www.elsevier.com/locate/procedia

Emerging Markets Queries in Finance and Business

Considerations on the correlation between Student test statistics in cases of simple and multiple linear regressions

Florin Marius Pavelescu[a,*]

[a]*Institute of National Economy, 13, Calea 13 Septembrie, Sector 5, Bucharest, Romania*

Abstract

The paper brings arguments in favour of the idea that Student test statistics may not be considered only in comparison with the critical (tabelled) values in order to determine the significance of estimated parameters. Interesting information on the quality of estimation can be obtained by studying the correlation between the Student test statistics related to each explanatory variable in case of simple and multiple regressions, respectively. This way, it is possible to identify the factors which determine the change the Student test statistics related to considered explanatory variable as the number of the other explanatory variables increase in the linear regression equation. In this context, a factorial model for the analysis of the above-mentioned change is proposed. The proposed analysis methodology is practically applied to a private consumption function estimated for Romania during the period 1990-2009.

Selection and peer-review under responsibility of Asociatia Grupul Roman de Cercetari in Finante Corporatiste

Keywords: Coefficient of collinear refraction, VIF (variance of inflation factor), transformed form of Student test statistics, standard error, consumption function, harmful collinearity

* Corresponding author. Tel.: 4-021-318-2467, fax: 4-021-318-2467:.
*E-mail address:*pavelescu.florin@yahoo.com;

Selection and peer-review under responsibility of Asociatia Grupul Roman de Cercetari in Finante Corporatiste
doi:10.1016/S2212-5671(15)01366-0

1. Introduction

Student test is frequently used in order to determine the significance of estimated parameters of linear regressions. But, in many cases, the Student test statistics of multiple linear regressions are sensibly different from the statistics obtained for the analyzed explanatory variables in case of simple linear regressions. The main cause of the respective differentiation is due to the collinearity, in other words because of departure from orthogonality between the explanatory variables (Glauber, Farrar, 1967). But if we consider some indicators such as Variance of Inflation Factor or coefficients of collinear refraction, we are able to explain the correlation between the Student statistics obtained in case of simple and multiple linear regressions.

2. Algebraically properties of Student test statistics in case of simple and multiple linear regressions

If we have in view that the standard Student test statistics is defined as the ratio between the estimated value and the standard error of the considered parameter and if ordinary least square method is used, in case of a simple linear regression, the Student test statistics related to explanatory variable x_k (t_{b1k}) is equal with:

$$t_{1k} = (m-2) \cdot \frac{R(x_k; y)}{\sqrt{1-R^2(x_k; y)}} \qquad (1),$$ where:

m= number of the observations

$R(x_k; y)$ = Pearson coefficient of correlation between the dependent variable y and explanatory variable x_k

Formula (1) is obtained because the estimated value of parameter related to considered explanatory variable (b_{1k}) can be computed with the help of formula:

$$b_{1k} = \frac{D(y)}{D(x_k)} \cdot R(x_k; y) \qquad (2),$$

where:

D(y)= standard deviation of the observed values of the dependent variable y

$D(x_k)$= standard deviation of the observed values of the explanatory variable x_k

Standard error of parameter b_{1k} ($SE(b_{1k})$) may be computed in case of a simple regression with the formula:

$$SE(b_{1k}) = \sqrt{\frac{1}{m-2}} \cdot \left(\frac{D(y)}{D(x_k)}\right) \cdot \sqrt{1-R^2(x_k; y)} \qquad (3)$$

If we deal with a multiple linear regression i.e. $y = a_n + \sum_{k=1}^{n} b_{nk} \cdot x_k$, the estimated parameter b_{nk} can be written as: $b_{nk} = b_{1k} \cdot T_{nk}$ (4), where:

T_{nk}= coefficient of collinear refraction[1]

Coefficient of collinearity refraction (T_{nk}) may be written using the formula[2]:

$$T_{nk} = \frac{1 - p_{jkwmed} \cdot R^2_{(n-1)xk}}{1 - R^2_{(n-1)xk}} \quad (5)$$ where: p_{jkwmed}= weighted arithmetical mean of ratios p_{jk}

$$p_{jk} = \frac{R(x_j; y)}{R(x_j; x_k) \cdot R(x_k; y)} \quad (6),$$ where:

$R(x_j, y)$ = Pearson coefficient of correlation between explanatory variable x_j and dependent variable.

$R(x_j, x_k)$ = Pearson coefficient of correlation between explanatory variable x_j and explanatory variable x_k.

$R(x_k, y)$ = Pearson coefficient of correlation between explanatory variable x_k and dependent variable.

$R^2_{(n-1)xk}$= Coefficient of determination of linear regression $x_k = C_{(n-1)xk} + \sum_{j=1}^{n-1} c_{(n-1)xj} \cdot x_{j(j \neq k)}$

Having in view E. Uriel Jimenez (2013) and making some algebraically transformations, the standard error of the estimated parameter b_{nk} ($SE(b_{nk})$ may be computed with the formula:

$$SE(b_{nk}) = \sqrt{\frac{1}{m-n-1}} \cdot \left(\frac{D(y)}{D(x_k)}\right) \cdot \sqrt{\frac{1 - R^2_{ny}}{1 - R^2_{(n-1)xk}}}, \quad (7)$$ where:

R^2_{ny}= Coefficient of determination of linear regression $y = a_n + \sum_{k=1}^{n} b_{nk} \cdot x_k$

Consequently, the Student test statistics related to explanatory variable x_k in case of a linear regression with n explanatory variables is determinated by the formula[3]:

---

[1] **Coefficient of collinear refraction** as a modeling factor of the estimated value of a parameter related to an explanatory variable in conditions of a multiple linear regression was firstly identified in F.M. Pavelescu (1986) and was named "coefficient of alignment". In F. M. Pavelescu (2010) the respective factor is named "coefficient of alignment to collinearity hazard". But the respective names given to respective coefficient is too long and in a way unclear. In fact, the respective coefficient plays a role similar to index of refraction of light when waves of light pass through different medium. The coefficient of collinear refraction explains the "strange" values of estimated parameters obtained in case multiple linear regressions**. If the respective coefficient is negative at least in case of one of the estimated parameters of a linear regression we are faced with a harmful collinearity.**

[2] In F.M. Pavelescu (2014 b) it was shown that formula (5) can be rigorously demonstrated in case of regressions with two and three explanatory variables. The other estimations made informally by the author of the present paper reveal that the respective formula is right even if the number of explanatory variables is greater than three.

[3] On the other hand, if we consider the variance of the inflation factor (VIF) defined as:

$VIF = \frac{1}{1 - R^2_{(n-1)xk}}$, t_{bnk} may be written also as: $t_{bnk} = \sqrt{m-n-1} \cdot R(x_k; y) \cdot \sqrt{\frac{VIF}{1 - R^2_{ny}}} \cdot (1 - p_{jkwmed} \cdot R^2_{(n-1)xk})$

$$t_{bnk} = \sqrt{m-n-1} \cdot R(x_k; y) \cdot \sqrt{\frac{1-R^2_{(n-1)xk}}{1-R^2_{ny}}} \cdot T_{bnk} \quad (8)$$

Formula (8) reveals that the Student test statistics are influenced by coefficient of collinear refraction. Consequently, the standard form of Student test statistics makes not possible a rapid and direct identification of harmful collinearity. In these conditions in (F. M. Pavelescu, 2013) it was proposed the use of a Transformed form of Student Test statistics ($TFST_{bnk}$), defined by formula:

$$TFST_{bnk} = \sqrt{m-n-1} \cdot \left|R(x_k; y)\right| \cdot \sqrt{\frac{1-R^2_{(n-1)xk}}{1-R^2_{ny}}} \cdot T_{bnk} \quad (9)$$

It is to note the transformed form of Student test statistics may be also considered in case of a simple linear regression ($TFST_{b1k}$) by using the formula:

$$TFST_{b1k} = \sqrt{m-2} \cdot \left|R(x_k; y)\right| \cdot \sqrt{\frac{1}{1-R^2(x_k; y)}} \quad (10)$$

This way, it is possible to make comparisons between Student test statistics obtained in case of simple and multiple linear regressions, respectively.

It is to note that transformed form of Student test statistics ($TFST_{bnk}$) may be used only in the context of linear regressions. In F.M. Pavelescu (2013) it was stated that the use of the respective transformed form would lead to some changes in estimation methodology of linear regressions, namely by adding of a new stage. In other words, at the beginning of the estimation process, the parameters values are computed in the same time with the computation of the transformed form of Student test statistics. If there are negative values of $TFST_{bnk}$ for any parameter, we are able conclude that the harmful collinearity has occurred in the respective linear regression and the results obtained may be rejected. If all $TFST_{bnk}$ are positive, we may conclude that there is no harmful collinearity and it is possible to further use the standard methodology of estimation.

3. Identification of modeling factors of the correlations between the Student test statistics in cases of simple and multiple linear regressions

If we consider the formulae (9) and (10)) it is possible to write:

$$TFST_{bnk} = TFST_{b1k} \cdot \sqrt{\frac{m-n-1}{m-2}} \cdot \sqrt{\frac{(1-R^2_{(n-1)xk}) \cdot (1-R^2(x_k; y))}{1-R^2_{ny}}} \cdot T_{bnk} \quad (11)$$

Analogously, we also have: $t_{bnk} = t_{b1k} \cdot \sqrt{\frac{m-n-1}{m-2}} \cdot \sqrt{\frac{(1-R^2_{(n-1)xk}) \cdot (1-R^2(x_k; y))}{1-R^2_{ny}}} \cdot T_{bnk}$ (12)

Formula (11) reveals the fact the correlation between the Student test statistics in case of simple and multiple linear regressions is determinted by two major factor, namely: a) the inverse of index of standard error change of the estimated parameter ($ISE(b_{(1-n)k})$ and b) the coefficient of collinear refraction (T_{nk}).

Index of change of standard error can be viewed as a product between the index of impact of decrease of the freedoms degree as a result of adding of the other explanatory variables (Idfd) and the inverse of index of normalized error change of the estimated parameter (In $SE(b_{(1-n)k})$).

N.B. We define normalized error of an estimated parameter ($nSE(b_{nk})$ as: $nSE(b_{nk}) = \frac{SE(b_{nk})}{\sqrt{m-n-1}}$ (13)

We can notice that $Idfd = \sqrt{\frac{m-n-1}{m-2}} < 1$ (14)

In $SE(b_{(1-n)k}$ can be viewed as a product between two factors, respectively:

$$InSE(b_{(1-n)k}) = \sqrt{\frac{1-R^2_{(n-1)xk}}{1-R^2_{(n-1)y}}} \cdot \sqrt{\frac{(1-R^2_{(n-1)y}) \cdot (1-R^2(x_k;y)}{1-R^2_{ny}}}$$ (15), where:

$R^2_{(n-1)y}$= Coefficient of determination of linear regression $y = B_{(n-1)y} + \sum_{j=1}^{n-1} b_{(n-1)xj} \cdot x_{j(j\neq k)}$

If we note $G = \sqrt{\frac{1-R^2_{(n-1)xk}}{1-R^2_{(n-1)y}}}$ *(16)*

and $H = \sqrt{\frac{(1-R^2_{(n-1)y}) \cdot (1-R^2(x_k;y)}{1-R^2_{ny}}}$ *(17)*

, it is possible to express $InSE(b_{(n-1)k})$ as:

$$InSE(b_{(1-n)k}) = G \cdot H$$ *(18).*

The use of factor G has the advantage to permit a comparison between the coefficients of determination of linear regressions in which y and x_k, respectively, are regressed as against the other explanatory variables considered in the multiple linear regression. Under these conditions, it is possible to reveal the most important

premise of relative change of the standard error of considered parameter as a result of adding of the other explanatory variables (x_j j=1..(n-1)). If G>1, we may affirm that there are created premises for an "efficient" adding of the considered explanatory variables in linear regression from the point of view of standard error decrease.

The value of factor H gives information's on the increase feature of the coefficient of determination of linear regression related to dependent variable as a result of adding of the considered explanatory variable.

If we note: $\beta = \frac{R^2_{(n-1)y}}{R^2_{1y}}$ (19),

and

$$\gamma = \frac{R^2_{ny}}{R^2_{(n-1)y}}, \quad (20)$$

it is possible to demonstrate that:

$$H > 1 \text{ if } \gamma > (1 - R^2_{1y}) + \frac{1}{\beta} \quad (21)$$

In other words, a condition to obtain H>1, is that relative increase of the coefficient of determination generated by adding of the analyzed explanatory variable in the linear regression equation is greater than the ratio between $R^2_{(n-1)y}$ and $R^2_{1y.}$

Consequently, we are faced with six possible situations regarding In $SE(b_{(1-n)k})$, namely:

I) G<1, H<1 and In $SE(b_{(1-n)k})$<1; II) G<1, H>1 and In $SE(b_{(1-n)k})$<1, III) G<1, H>1 and In $SE(b_{(1-n)k})$>1

IV) G>1, H<1 and In $SE(b_{(1-n)k})$<1; V) G>1, H<1 and In $SE(b_{(1-n)k})$>1, VI) G>1, H>1 and In $SE(b_{(1-n)k})$>1

Regarding to coefficient of collinear refraction, we may note that, depending on the interactions between the Pearson coefficients of correlation, the respective indicator may take values in different intervals. But as a rule, the above-mentioned indicator values are as polarized as the related VIF is greater. On the other hand, as the number of explanatory variables increases, the values of coefficients of collinear refraction show a tendency to diminish. Therefore, as a tendency, Student test statistics related to considered explanatory variable in case of a simple linear regression is greater than the respective statistics in case of a multiple linear regression.

4. A numerical example. Estimation of a Consumption Function for Romania during the period 1990-2009

In order to illustrate the proposed methodology for identification of contribution of modeling factors to the correlation between the Student test statistics in case of simple and multiple linear regression, respectively we have estimated the parameters of a consumption function for Romania during the period 1990-2009.

The form of proposed consumption function is:

$$CH_t = a_3 + b_3 \cdot CH_{(t-1)} + c_3 \cdot BEC_t + d_3 \cdot M_t \quad (22),$$

where:

CH_t = private consumption in year t
BEC_t= state budget expenditures in year t
M_t = imports in year t

The estimation of the consumption function parameters gave the following result:

$CH_t = -0.1707 + 0.7353*CH_{t-1}-0.1616*BEC_t+0.9113*M_t$ $R^2_{3CHt}= 0.9833$
(-0.0263) (5.0245) (-0.4424) (4.6731)

N.B. *In the brackets there presented the standard Student test statistics*
R^2_{3CHt}= coefficient of determination of the estimated linear regression equation

The estimation of simple linear regressions for each of the explanatory variables led to the following results:

$CH_t = -0.6588 + 1.0467*CH_{t-1}$ $R^2_{1CHt/CH(t-1)}= 0.9333$ $R_{1CHt/CH(t-1)}= 0.9661$
(-0.1316) (15.8713)

$CH_t = -22.6159 +3.1278*BEC_t$ $R^2_{1CHt/BECt}= 0.5496$ $R_{1CHt/BECt}= 0.7413$
(-1.0643) (4.6864)

$CH_t = 9.2181 +2.2171*M_t$ $R^2_{3CHt/Mt}= 0.8661$ $R_{3CHt/Mt}= 0.9306$
(1.4250) (10.7903)

Because all the Pearson coefficients of correlation between dependent variable and explanatory variables are positive, the transformed Student test statistics are the same with the standard form of the respective statistics.

Therefore, it is possible to determine the coefficients of collinear refraction, respectively:

$T_{3CH(t-1)}$=0.7025, T_{3BECt}=-0.0517, T_{3Mt} = 0.4110

It is to note that in estimated private consumption function there is a harmful collinearity which is manifest in case of explanatory variable BEC_t. Estimation of private consumption function with only two explanatory variables gave the following results:

$CH_t = 19.9722 + 1.3416*CH_{t-1}-1.3424*BEC_t$ $R^2_{2CHt/CH(t-1)BECt}= 0.9604$
(2.7624) (13.2863) (-3.4138)

$CH_t = -24.3658 + 0.1616*BEC_t+1.7800*M_t$ $R^2_{2CHt/BECtMt}= 0.9569$
(-3.6004) (5.9813) (12.6702)

$CH_t = -2.7843+ 0.6771*CH_{t-1}+0.9709*M_t$ $R^2_{3CHt/CH(t-1)Mt}= 0.9831$
(-1.0653) (10.8353) (7.0671)

Estimation of the linear regressions in which dependent variable is each of the explanatory variables of initial proposed private consumption function, further permitting to determine VIF led to the following results :

$M_t =22.1039 + 0.6654*CH_{t-1}-1.2958*BEC_t$ $R^2_{2Mt}= 0.8439$
(3.6673) (7.9038) (-3.9528)

$CH_{t--1} = -32.9047 + 2.2437*BEC_t+1.1814*M_t$ $R^2_{2CH(t-1)}= 0.9427$
(-4.5698) (8.4751) (7.9038)

$BEC_t = 16.1775 - 0.3654*Mt_{t-1}+0.3604*CH_{(t-1)}$ $R^2_{3CHt/CH(t-1)Mt}= 0.8604$
(-1.0653) (-3.9528) (7.9038)

In these conditions, we are able to compute the modeling factors of the inverse of index of standard error change of the estimated parameter ($ISE(b_{(1-n)k})$), namely: Idfd, G and H, which are presented in table 1.

Table 1. Modeling factors of inverse of index of change of standard error change of estimated parameters of Romania's private consumption function during period 1990-2009

Explanatory variable	Idfd	G	H	$InSE(b_{(1-3)})$	$ISE(b_{(1-3)})$
CH_{t-1}	0.9428	1.2036	0.3971	0.4780	0.4506
BEC_t	0.9428	1.7992	1.0776	1.9387	1.8279
M_t	0.9428	3.0355	0.3682	1.1175	1.0536

We may remark that G>1 in case of all the three explanatory variables, creating premises for diminishing of the standard error of estimated parameters. Factor F is greater than 1 only in case of explanatory variable BEC_t. Consequently, as a result of adding of the other explanatory variables, normalized standard error increases in case of CH_{t-1} and decreases in case of BEC_t and M_t.

The size of $ISE(b_{(1-3)})$ related to explanatory variables are inversely correlated with the size of coefficients of collinear refraction (T_{3k}). Therefore, it is possible that harmful collinearity occurs in the same time with a sensible decrease of the standard error of estimated parameter as in case of explanatory variable BEC_t (table 2).

Table 2. Modeling factors of Student test statistics related to explanatory variables of Romania's consumption function during period 1990-2009

Explanatory variable	t_{1k}	$ISE(b_{(1-3)})$	T_{3k}	t_{3k}
CH_{t-1}	15.8713	0.4506	0.7025	5.0245
BEC_t	4.6864	1.8279	-0.0517	-0.4424
M_t	10.7903	1.0536	0.4110	4.6731

On the other hand, we may note that $ISE*T_{3k}<1$ for all the three considered explanatory variables. Consequently, we are in a situation when the Student test statistics of simple linear regressions are smaller than the respective statistics determinate in case of multiple regression.

It is also important to remark that Student test statistics of explanatory variables of multiple linear regression (t_{3k}) which was computed by considering the product of Student test statistics of simple linear regression (t_{1k}), inverse of index of standard error change of the estimated parameter ($ISE(b_{(1-3)k})$) and coefficients of collinear refraction (T_{3k}) are the same with the results obtained when the considered multiple linear regression was initially run.

5. Conclusions

Theoretical considerations and practical example presented in this paper permit to identify the modeling factors of Student test statistics in case of multiple linear regression an to explain the occurrence of eventual "strange" values in comparison with expected one, having by in view the absolute value of Pearson coefficient of correlation between dependent variable and considered explanatory variable. Therefore, it is possible to emphasize the determinant impact of variance inflation factor and of the coefficient of collinear refraction on the correlation between the Student test statistics in simple and multiple linear regression case, respectively.

The above-mentioned correlation permits to give an image of the factors which determine the evolution of standard error of the estimated parameter as the number of explanatory variables increases. On the other hand,

the numerical example presented in this paper shows that the decrease of standard error of estimated parameter is not a sufficient condition in order to avoid the harmful collinearity occurrence.

Identification of modeling factors permit to compare the Student test statistics related to analyzed explanatory variable in the context of change of number of explanatory variables considered in linear regression equation. As the number of explanatory variables increases, the Student test statistics tend to diminish. Consequently, it is more and more difficult to validate as been significant all the explanatory variables if their number in the linear regression is bigger and bigger.

References

A. Alin – Multicollinearity. Focus Article, Willey Interdisciplinary Reviews: Computational Statistics, Vol. 2, Issue 3, 2010

D Belsey – Multicollinearity: Diagnosing its presence and assessing the potential damage it causes least square estimation, *NBER Working Paper no. 154, Cambridge, Massachusetts, October. 1976*

D. Belsey - Conditional diagnostics: collinearity and weak data in regression, Wiley Series in Probability John Wiley, New York., 1991

C. Conrad -Applied Regression Analysis. Lectures Notes, Department of Economics, Pomona College, Claremont, California, Spring, 2006.

C.Dougherty - Introduction to econometrics-Second Edition, Oxford University Press, 2002.

D. Farrar, R. Glauber, - Multicollinearity in regression analysis: The problem revisited, *Review of Economics and Statistics, February.* 1967

W. Greene (1993)-Econometric Analysis, Mc Millan Publishing Company, New York.

A. Isaic - Maniu, C. Mitruţ, V. Voineagu (1996)- Statistica pentru managementul afacerilor, Editura Economică, Bucureşti, 1996

D. Jula (2003) - Introducere în Econometrie, Ed Professional Consulting, Bucureşti.

F. M. Pavelescu - Some Considerations regarding the Significance of Cobb-Douglas Production Function. A new Aprproach., Revue Romaine des Sciences Sociales, Tome 30, nr. 1-2/1986

F.M: Pavelescu- Impact of collinearity on the estimated parameters and classical statistical tests values of multifactorial linear regressions in conditions of O.L.S. în Romanian Journal of Economic Forecast no.2/2005

F.M. Pavelescu (2014a)- Methodological considerations regarding the estimated returns to scale in case of Cobb Douglas production function, Elsevier Procedia Economics and Finance, vol.8/2014, pag. 535-542, FINE 678 /12 april 2014

F.M. Pavelescu (2014b)- An extension of the methodology of using the Student Test in case of a linear regression with three explanatory variables, Romanian Economic Journal no. 1/2014.

E.Uriel Jimenez- Introduction to Econometrics. Electronic textbook, University of Valencia, version 09-2013 (www. uv. es/ uriel/libroin. htm, accessed 14-th october 2014).

R.Wiliams - Sociology Graduate Statistics, University of Notre Dame, Indiana.2008