ISSN 1991- 8690 الترقيم الدولي ٨٦٩٠ - ١٩٩١
website: http://jsci.utq.edu.iq Email: utjsci@utq.edu.iq

Correction for multicollinearity between the explanatory variables to estimation by using the Principal component method

Ali L .Areef

Science college – Thi- Qar University

Abstract

In the most of applications of regession ,no explanatory orthogonal will exist,but it is connected too strongly to the extennt that the results would be far from being exquisite. So its too difficult to expect the effects upon the individual variabls within the range of regression equation .Also the values estimated here concerning the factors could be slight in data. The non-orthogonality is said to be the problem of multicolinearity going side by side with the factors of unstable,estimated regression. This case,however,comes out from the strong linear relation between explanatory variants.To solve this problem, a method of the pricipal components is used;that which depends upon the fact that each linear type mihgt be reformulated as to the group of the orthogonal,explanatory variables; these in turn can be obtained as linearstructures for the orthogonal (basic) explanatory variables through the Barr'let norm,as a test formula, as away to know whetherthe roots possess a sufficient quality for a linear relation As for the practical side,or applied of this research concerning the special data of the consumption of the individual in USA as dependent variable and wage income,non wage-non farm income,farm income are explanatory variables.the characters rootindicated to the collinearity ; the result is that the four variables can be treated as two factors only ,a ststistical programme is here used that is (SPSS,Minitab) for the analysis of the data.

Introduction

A regression model that involves more than repressor variable is called a multiple regression model. In other words it is a linear relationship between a dependent variable (Y) and two or more independent (explanatory) (X_i) therefore used as approximating functions. That is true functional relationship between Y and $X_1, X_2, \ldots\ldots, X_k$ is unknown, but over

certain ranges of the repressors variables, the linear regression model is an adequate approximation to the true unknown function .If the model fits the data will, The R^2 value will be high ,and the corresponding P value will be low (P value is the observed significance level at which the null hypothesis is rejected). in addition , the multiple regressions also report an individual P value for each independent variable .A low P value here means that this particular independent variable significantly improves the fit of the model .It is calculated by comparing the goodness-of-fit of the entire model to the goodness-of-fit when that independent variable is omitted . If the fit is much worse when that variable is omitted from the model, the P value will be low, telling that the variable has a significant impact on the model.[1]

Consider the following multiple regression models

$$Y = X\beta + \varepsilon \quad \ldots\ldots\ldots \quad (1)$$

Where Y is an (n×1) vector of responses,X is an (n×p)matrix of the regressor variables,β is (p×1) vector of unknown constant; and ε is an (n×1) vector of random errors,with $\varepsilon_i \approx N(0,\sigma^2)$ .it will be convenient to assume that the regressor variables are standardized.consequently , $X^{'}X$ is (p×p)matrix of correlations between the regessors and $X^{'}y$ is (p×1) vector of correlation between the regessors and the response. Let the j^{th} column of X matrix be denoted by X_j ,so that $X = \lfloor X_1, X_2, ..., X_p \rfloor$ .Thus X_j contains the n level of the regressor variable.formally multicollinearity can be defined as the linear dependence of the columns of X.the vectors are linearly dependent if there is a set of constants $\lambda_1, \lambda_2, ..., \lambda_k$, not all zero such that

$$\sum_{j=1}^{k} \lambda_j X_j = 0 \quad \ldots\ldots\ldots \quad (2)$$

If equation (2) holds exactly for a subset of the columns of X,then the rank of the $X^{'}X$ matrix is less than p and $(X^{'}X)^{-1}$ does not exist [4]

Multicollinearity

The term Multicollinearity refers to a situation in which there is an exact (or nearly exact) linear relation among two or more of the input variables, exact relations usually arise by mistake or lack of understanding.

If the goal is simply to predict Y for a set of X variables, the multicollinearity is not a problem. The predictions will still be accurate, and the overall R^2 (or adjusted $R^{2)}$ quantifies how will the model predicts the Y values. However, if the goal is to understand how the various effect Y, then multicollinearity is a big problem? One problem is that the individual P values can be misleading (a P value can be high, even though the variable is important) .The second problem is that the confidence intervals on the regression coefficient will be very wide. The confidence intervals may even include zero, which means one cannot even be confident whether an increase in the X value is associated with in increase, or a decrease, in Y. Because the confidence intervals are so wide ,excluding a subject (or adding a new one) can change the coefficients dramatically and may even change their signs.

There are four primary sources of multicollinearity

1- The data collection method employed
2- Constraints on the model or in the population.
3- Model specification
4- An over defined model

The data collection method can lead to multicollinearity problems when the analyst samples only a subspace of the region of the repressors defined in equation (2) [7]

The method of principal components

The aim of the method of the principal components is the construction out of a set of variable, X_j's, j=1, 2,……, k of new variables Z_i called principal components, which are linear combinations of the X_j's.

$$Z_{1t} = a_{11}X_{1t} + a_{21}X_{2t} + ... + a_{k1}X_{kt} \quad , \ t=1, 2\ldots n$$

$$Z_{2t} = a_{12}X_{1t} + a_{22}X_{2t} + ... + a_{k2}X_{kt}$$

.

.

$$Z_{kt} = a_{1k}X_{1t} + a_{2k}X_{2t} + ... + a_{kk}X_{kt} \quad \ldots\ldots\ldots \quad (3)$$

Here (a's) are called (loadings) which principal components are chosen, so constructed principal components satisfy two conditions; [3]

1- The principal components are uncorrelated.

2- The first principal component Z_1 absorbs and accounts for the maximum

Possible proportion of the total variation in the set of all X_j's,

The second principal component absorbs the maximum of the remaining variation in the X_j's (after allowing for the variation accounted for the first principal component) and so on.

Test for the significance of the loadings

The loadings are in fact similar to correlation coefficients. This test does not take into account the number of variables, X's in the set ,and the order of extraction of the principal components .Burt and Banks [1947] have suggested the following adjustment to the standard error of the correlation coefficient in order to obtain the standard errors of the loadings [2]

$$s(a_{ij}) = [s(r_{xj}.{}_{xm})](k/k+l-i)^{1/2} \quad \ldots\ldots.(4)$$

Where

K=number of X's in the set

L=subscript of Z, i.e. the order of its extraction (the position of Z in the extraction Process). The .Burt and Banks fomula, clearly takes into account both the number of X's and the order of extraction of Z's

Barlett's Criteria for the number of principal components to be extracted

Assume the latent roots are computed for k variables $\lambda_1, \lambda_2 \ldots \lambda_k$ and the first r roots $\lambda_1, \lambda_2 \ldots \lambda_r$ (for r<k) seem both sufficiently large and sufficiently different to be retained. The question then whether the remaining (k-r) roots are sufficiently alike for one to conclude that the associated Z's should be retained in the analysis .Bartlett [1954] has suggested that the quantity

$$\chi_c^{\,2} = nl_n[(\lambda_{r+1}, \lambda_{r+2}, ..., \lambda_{r+k})^{-1}(\lambda_{r+1}, \lambda_{r+2}, ...\lambda_k/(k-r))^{k-r}] \quad \ldots\ldots.(5)$$

Has χ^2-distribution (approximately) with $v = 1/2(k-r-1)(k-r-2)$ degrees of freedoms. The null hypothesis has assumed equality of the excluded latent roots, i.e.

$$H_0 = \lambda_{r+1} = \lambda_{r+2} = ... = \lambda_r \quad \ldots\ldots\ldots.(6)$$

If $\chi_c^{\,2} > \chi_{(1-\alpha,v)}$ we reject the null hypothesis, that is we accept that the excluded latent roots are not equal; hence, we should include additional Z's in our analysis [6]

Principal component regression

Let the model under consideration be,

$Y = X\beta + \varepsilon$

Let $X'X = T\Lambda T'$, where $\Lambda = diag(\lambda_1, \lambda_2, ...\lambda_k)$ ……..(7)

A P × P diagonal matrix of the eigenvalues is $X'X$

In addition, T is a P × P orthogonal matrix whose columns are the eigenvectors associated with $\lambda_1, \lambda_2 \ldots \lambda_k$. Then the above model can be written as

$Y = XTT'\beta + \varepsilon$, $TT' = I$, Is the identity matrix

$= Z\alpha + \varepsilon$, Where

$Z = XT$, $\alpha = T'\beta$ …….(8)

Where $T = (a_1, a_2, \ldots, a_3)$

$Z'Z = T'X'XT = T'T\Lambda T'T = \Lambda$ ……(10)

The columns of Z, which define a new set of orthogonal repressors

, such as $Z = [Z_1, Z_2, \ldots\ldots Z_3]$ are referred to as principal components [5]

The least square estimator of α is

$\hat{\alpha} = (Z'Z)^{-1}Z'Y = \Lambda^{-1}Z'Y$ ……..(11)

And the covariance matrix of $\hat{\alpha}$ is

$V(\hat{\alpha}) = \sigma^2(Z'Z)^{-1} = \sigma^2\Lambda^{-1}$ …(12)

Thus a small eigenvalues of $X'X$ means that the variance of the corresponding regression coefficient will be large. Since $Z'Z = \sum_{i=1}^{k}\sum_{j=1}^{k} Z_i Z_j' = \Lambda$. We often refer to the eigenvalues λ_j as the variance of the jth principal component. If all λ_j equal to unity, the original regressors are orthogonal, while if a λ_j is exactly equal to zero, this implies a perfect linear relationship between the original regressors. One or more λ_j near to zero implies that multicollinearity is present.

The principal component regression approach combats multicollinearity by using less than the full set of principal components in the model. To obtain the principal components estimator, assume that the regressors are arranged in order of decreasing eigenvalues, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k > 0$ suppose that the last of these eigenvalues is approximately equal to zero.

In principal component regression the principal component corresponding to near zero eigenvalues are removed from the analysis and least squares applied to the remaining components. That is

$\hat{\beta} = T\hat{\alpha}$ ….(13)

Where $a_1 = a_2 = \ldots = a_{k-s} = 1$ and $a_{k-s+1} = a_{k-s+2} = \ldots = a_k = 0$

Thus the principal components estimator is

$\hat{\beta} = [\hat{\alpha_1}\ \hat{\alpha}_2\ \ldots\ldots \hat{\alpha}_{k-s}\ \ldots . 0\ 0 \ldots 0]'$ ……(14)

Application

The data are represented to the United States economy, where Y=consumption, X_1 =wage income, X_2 =non-wage, non farm income, X_3 =farm income.

Table 1:

Y	X_1	X_2	X_3
62.8	43.41	17.10	3.96
65.0	46.44	18.65	5.48
63.9	44.35	17.09	4.37
67.5	47.82	19.28	4.51
71.3	51.02	23.24	4.88
76.6	58.71	28.11	6.37
86.3	87.69	30.29	8.96
95.7	76.73	28.26	9.76
98.3	75.91	27.91	9.31
100.3	77.62	32.30	9.85
103.2	78.02	31.39	7.21
108.9	83.57	35.61	7.39
108.5	90.59	37.58	7.98
111.4	95.47	35.17	7.42

Table (2) Model Summary

Change Statistics					Std. Error of the Estimate	Adjusted R Square	R Square	R	Model
Sig. F Change	df2	df1	F Change	R Square Change					
.000	10	3	37.683	.919	6.05969	.894	.919	.959(a)	1

a Predictors: (Constant), X3, X2, X1

a Dependent Variable: Y

Table(3) KMO and Bartlett's Test

.833	Kaiser-Meyer-Olkin Measure of Sampling Ade quacy.	
62.868	Approx. Chi-Square	Bartlett's Test of Sphericity
6	df	
.000	Sig.	

Table(4) **Coefficients(a)**

Sig.	t	Standardized Coefficients	Unstandardized Coefficients			Model
		Beta	Std. Error	B		
.021	2.732		6.845	18.703	(Constant)	1
.251	1.219	.385	.312	.380	X1	
.077	1.969	.539	.720	1.418	X2	
.711	.381	.059	1.400	.533	X3	

a Dependent Variable: Y

we see that coefficient of determination R^2 =91.9 % is highly significant but in contrast all the β's are insignificant .the computation reveal that the cause of multicollinearity lies mainly in the Intercorrelation between X_1 and X_2 .now it has been proved that the data are highly multicollinear.

The principal component analysis based correlation method yields the summary table (5) .

Table (5)

principal Component Analysis

Eigenanalysis of the Correlation Matrix

Eigenvalue Var(λ_1)	2.6636	0.2880	0.0484
Proportion	0.888	0.096	0.016
Cumulative %of variation	0.888	0.984	1.000

Variable	PC1	PC2	PC3
X1	-0.598	0.253	0.760
X2	-0.583	0.514	-0.630
X3	-0.550	-0.820	-0.160

The three principal component are:

$Z_1 = -0.598X'_1 - 0.583X'_2 - 0.550X'_3$

$Z_2 = 0.253X'_1 + 0.514X'_2 - 0.820X'_3$

$Z_3 = 0.760X'_1 - 0.630X'_2 - 0.160X'_3$

If we retain all of these three component we will get the estimate similar to the OLS estimates. table (5) shows the coefficient of correlation between the first principal Z_1 and (x_1, x_2, x_3) are quite large , also the correlation between Z_2 and (x_1, x_2, x_3)

But the relationships between Z_3 and (x_1, x_2, x_3) are not very strong . it means that the first two principal components are sufficient to describe the maximum variation in X's we see that all the coefficints (loadings) of the first principal component are significant.only

third co-efficint of the second principal component is significant and not even a single co-efficient of the third pricipal component is significant .

Conclusions

1- The principal components and there co-efficint (loadings) are obtained by using the correlation matrix of the regressors.All the loading of the third pricipal component are insignificant,moreover,the correlation between orginal variable $\dot{X}$s (standardized) and the third pricipal component are insignificant .we therefore,exclude the third pricipal component from the analysis and retain only the first two components.
2- We observe that the principal components regression technique provides the best estimates of the coefficients of the population regression function ,in particular when the sample data are suffering from multicollinearity .
3- If the original variables are uncorrelated then there is no use of the principal components analysis.multicollinearity ,if present among the repressors, seriously affect the property of minimum variance of the OLS estimates.
4- If the purpose is just of the forecasting or prediction, then the existence of multicollinearity dose not harm any more but if aim is to get the precise estimates.

References

1- Butler,N.A.Denham, M.C.(2000) " The peculiar shrinkage properties of partial least Squares regression ",J.R.Statist.Soc,B62(3), 585-593.

2- Burt, C. and Banks, C. (1947) "A factor analysis of body measurement for British adult males", Ann. Eugene,13,238-256.
3 -Draper,N.R,Smith,H.(2003)"Applied regression analysis"3th edition,Wiley,New York
4 -Fareebrother (1972) "Principal component estimators and minimum mean squares error criteria in regression analysis ",Rev. of Econ.and statistics,54,332-336 .
5- Fomby,T.B. and Hill,R.c .(1978) "Multicollinearity and the value of a priori information ",,comm.. in statist ,A 8,477-486.
6- Helland , I. and Almoy,T.(1994) "Comparison methods when only a few Components are relevant ",JASA 89,583-591.
7- Samprit,Ch and Bertram,P.translated by Mohammad .M,(1989) "Regression analysis by example" p181-214.

الخلاصة

في أكثر تطبيقات الانحدار لا تكون المتغيرات التفسيرية متعامدة . بل مرتبطة بقوة إلى الحد الذي تكون فيه النتائج غير واضحة لذا فمن الصعوبة تقدير التأثيرات على المتغيرات المنفردة في معادلة الانحدار وان القيم التقديرية للمعاملات تكون حساسة جدا للتغيرات الطفيفة في البيانات. إن حالة عدم التعامد يقال لها مشكلة التعدد الخطي والذي يترافق مع معاملات الانحدار التقديرية غير المستقرة .وهذه الحالة تنتج من وجود علاقة خطية قوية بين المتغيرات التفسيرية. ولحل هذه المشكلة تم استخدام طريقة المركبات الأساسية التي تعتمد على حقيقة أن كل نموذج خطي يمكن إعادة صياغته بدلالة مجموعة من المتغيرات التفسيرية المتعامدة هذه المتغيرات يتم الحصول عليها كتراكيب خطية للمتغيرات التفسيرية الأصلية مع التطرق إلى معيار بار ليت كصيغة اختبار لكيفية معرفة فيما إذا كانت الجذور تمتلك صفة(الكفاية) للعلاقة الخطية. تناول الجانب التطبيقي من البحث البيانات الخاصة باستهلاك الفرد في الولايات المتحدة كمتغير معتمد ودخل الفرد (الأجر) ولادخل ولا اجر من المزرعة ودخل من المزرعة كمتغيرات مستقلة .إن ظهور جذر مميز صغير يشير إلى التعدد الخطي وتم التوصل إلى انه يمكن التعبير عن المتغيرات الأربعة بعاملين فقط . تم استعمال البرنامج الإحصائي(SPSS,Minitab)في عملية تحليل البيانات.