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 regression, no explanatory orthogonal will exist, but it is connected too strongly to the extent that the results would be far from being exquisite. So its too difficult to expect the effects upon the individual variables 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 principal components is used; that which depends upon the fact that each linear type might be reformulated as to the group of the orthogonal, explanatory variables; these in turn can be obtained as linear structures for the orthogonal (basic) explanatory variables through the Barr'let norm, as a test formula, as away to know whether the 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 root indicated to the collinearity; the result is that the four variables can be treated as two factors only, a statistical 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, ...., 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

\[ Y = X\beta + \varepsilon \]

……… (1)

Where Y is an (n×1) vector of responses, X is an (n×p) matrix of the regressor variables, \( \beta \) is (p×1) vector of unknown constant; and \( \varepsilon \) is an (n×1) vector of random errors, with \( \varepsilon \sim N(0, \sigma^2) \). It will be convenient to assume that the regressor variables are standardized. Consequently, \( X'X \) is (p×p) matrix of correlations between the regressors and \( X'y \) is (p×1) vector of correlation between the regressors and the response. Let the \( j^{th} \) column of X matrix be denoted by \( X_j \), so that \( X = [X_1, X_2, \ldots, X_p] \). 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 \( \lambda_1, \lambda_2, \ldots, \lambda_k \), not all zero such that

\[ \sum_{j=1}^{k} \lambda_j X_j = 0 \]

………. (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 well the model predicts the Y values. However, if the goal is to understand how the various effects 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 an 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, $j=1,2,\ldots,k$ of new variables $Z_i$ called principal components, which are linear combinations of the $X_j$'s.

\[
Z_{it} = a_{i1}X_{it} + a_{i2}X_{2t} + \ldots + a_{ik}X_{kt}, \quad t=1,2\ldots n
\]

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]

\[
s(a_{ij}) = [s(r_{ij,sm})](k/k + l - i)^{1/2}\]

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 formula, 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. Then the question becomes 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 the quantity

\[
\chi^2_c = n[l_n(\lambda_{r+1}, \lambda_{r+2}, \ldots, \lambda_{r+k})^{-1}(\lambda_{r+1}, \lambda_{r+2}, \ldots, \lambda_k (k-r))^{k-r}]
\]

Has $\chi^2$-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.

$H_0 = \lambda_{r+1}= \lambda_{r+2} = \ldots = \lambda_r$

If $\chi^2_c > \chi^2_{(v-\alpha,\nu)}$ 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 \( \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k) \) ............(7)

A P \times P diagonal matrix of the eigenvalues is \( X'X \)

In addition, \( T \) is a P \times 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

\[ Y = XTT'\beta + \varepsilon, \quad TT' = I \], Is the identity matrix

\[ Z = XT \]

Where \( T = (a_1, a_2, \ldots, a_k) \)

\[ ZZ = T'X'X = T'T = \Lambda \] ...................(10)

The columns of \( Z \), which define a new set of orthogonal repressors such as \( Z = [Z_1, Z_2, \ldots, Z_3] \) are referred to as principal components [5]

The least square estimator of \( \alpha \) is

\[ \hat{\alpha} = (Z'Z)^{-1}ZY = \Lambda^{-1}ZY \] ............(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 \( Z'Z = \sum_{i=1}^{k} \sum_{j=1}^{k} Z_iZ_j' = \Lambda \). We often refer to the eigenvalues \( \lambda_j \) as the variance of the jth principal component. If all \( \lambda_j \) equal to unity, the original repressors are orthogonal, while if a \( \lambda_j \) is exactly equal to zero, this implies a perfect linear relationship between the original repressors. One or more \( \lambda_j \) 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 repressors 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

\[ \hat{\beta} = [\alpha_1^\wedge \alpha_2^\wedge \ldots \alpha_{k-s}^\wedge \ldots 0 0 \ldots 0]^\prime \] .......(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:

<table>
<thead>
<tr>
<th>Y</th>
<th>X_1</th>
<th>X_2</th>
<th>X_3</th>
<th>X_4</th>
</tr>
</thead>
<tbody>
<tr>
<td>62.8</td>
<td>43.41</td>
<td>17.10</td>
<td>3.96</td>
<td></td>
</tr>
<tr>
<td>65.0</td>
<td>46.44</td>
<td>18.65</td>
<td>5.48</td>
<td></td>
</tr>
<tr>
<td>63.9</td>
<td>44.35</td>
<td>17.09</td>
<td>4.37</td>
<td></td>
</tr>
<tr>
<td>67.5</td>
<td>47.82</td>
<td>19.28</td>
<td>4.51</td>
<td></td>
</tr>
<tr>
<td>71.3</td>
<td>51.02</td>
<td>23.24</td>
<td>4.88</td>
<td></td>
</tr>
<tr>
<td>76.6</td>
<td>58.71</td>
<td>28.11</td>
<td>6.37</td>
<td></td>
</tr>
<tr>
<td>86.3</td>
<td>87.69</td>
<td>30.29</td>
<td>8.96</td>
<td></td>
</tr>
<tr>
<td>95.7</td>
<td>76.73</td>
<td>28.26</td>
<td>9.76</td>
<td></td>
</tr>
<tr>
<td>98.3</td>
<td>75.91</td>
<td>27.91</td>
<td>9.31</td>
<td></td>
</tr>
<tr>
<td>100.3</td>
<td>77.62</td>
<td>32.30</td>
<td>9.85</td>
<td></td>
</tr>
<tr>
<td>103.2</td>
<td>78.02</td>
<td>31.39</td>
<td>7.21</td>
<td></td>
</tr>
<tr>
<td>108.9</td>
<td>83.57</td>
<td>35.61</td>
<td>7.39</td>
<td></td>
</tr>
<tr>
<td>108.5</td>
<td>90.59</td>
<td>37.59</td>
<td>7.98</td>
<td></td>
</tr>
<tr>
<td>111.4</td>
<td>95.47</td>
<td>35.17</td>
<td>7.42</td>
<td></td>
</tr>
</tbody>
</table>

Table (2) Model Summary

<table>
<thead>
<tr>
<th>Sig. F Change</th>
<th>df2</th>
<th>df1</th>
<th>F Change</th>
<th>R Square Change</th>
<th>Std. Error of the Estimate</th>
<th>Adjusted R Square</th>
<th>R Square</th>
<th>R</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>.000</td>
<td>10</td>
<td>3</td>
<td>37.683</td>
<td>0.919</td>
<td>6.05958</td>
<td>.884</td>
<td>.919</td>
<td>.959(a)</td>
<td>1</td>
</tr>
</tbody>
</table>

a Predictors: (Constant), X3, X2, X1
a Dependent Variable: Y

Table (3) KMO and Bartlett’s Test

| .833 | Kaiser-Meyer-Olkin Measure of Sampling Adequacy. |
| 62.868 | Approx. Chi-Square | Bartlett’s Test of Sphericity |
| 6 | df |
| .000 | Sig. |
we see that coefficient of determination $R^2 = 91.9\%$ is highly significant but in contrast all the $\beta$'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**

<table>
<thead>
<tr>
<th>Eigenanalysis of the Correlation Matrix</th>
<th>2.6636</th>
<th>0.2880</th>
<th>0.0484</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion</td>
<td>0.888</td>
<td>0.096</td>
<td>0.016</td>
</tr>
<tr>
<td>Cumulative % of variation</td>
<td>0.888</td>
<td>0.984</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>-0.598</td>
<td>0.253</td>
<td>0.760</td>
</tr>
<tr>
<td>$X_2$</td>
<td>-0.583</td>
<td>0.514</td>
<td>-0.630</td>
</tr>
<tr>
<td>$X_3$</td>
<td>-0.550</td>
<td>-0.820</td>
<td>-0.160</td>
</tr>
</tbody>
</table>

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 coefficients (loadings) of the first principal component are significant. only
third co-efficient of the second principal component is significant and not even a single co-efficient of the third principal component is significant.

Conclusions
1- The principal components and there co-efficient (loadings) are obtained by using the correlation matrix of the regressors. All the loading of the third principal component are insignificant, moreover, the correlation between original variable X’s (standardized) and the third principal component are insignificant. We therefore, exclude the third principal 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 does not harm any more but if aim is to get the precise estimates.

References
4 -Fareebrother (1972)”Principal component estimators and minimum mean squares error criteria in regression analysis “,Rev. of Econ.and statistics,54,332-336.
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.