Available online at www.elixirpublishers.com (Elixir International Journal)

Statistics

Elixir Statistics 49 (2012) 10108-10112

ISSN: 2229-712X

The detention and correction of multicollinearity effects in a multiple regression diagnostics

Osemeke Reuben Friday and Chris Emenonye
College of Education, Agbor Department of Mathematics, Delta State, Nigeria.

ARTICLE INFO

Article history:
Received: 6 May 2012;
Received in revised form:
16 August 2012;
Accepted: 22 August 2012;

Keywords
Pair-wise correlation matrix;
Variance Inflation factor (VIF);
Interactive plots;
Regression diagnostics;
Transformation.

ABSTRACT

A dynamic graphical display among the set of independent variables (partial regression plot), tolerance value, variance inflation factor (VIF) and pair wise correlation matrix among the set of predictors offers a variety of measures for assessing the problem of colinearity and multicollinearity in a multiple regression diagnostics. Multicollinearity is a violation of one of the assumption of regression analysis. Many diagnostics measures have been proposed for detecting multicollinearity. A tolerance value of less than 0.10 or 0.20 which is equivalent to a VIF of 5 or 10, p value>0.05 and the pair-wise correlation showing a linear dependence of 0.90 and above. A transformation is carried out to remove the problem of multicollinearity and the removal will help to improve on the model (yi = $\beta0 + \beta\ _1X_1 + \beta_2X_2 + \ldots\ldots + \beta_kX_{k,}$ improve coefficient of determination (r^2) and validate any of the assumption of regression analysis of homocesdasticity, linearity, normality and independence of the observations. Examples using stimulated small data sets illustrate this approach

Introduction

Statement of problem

Multiple regressions is most effect at identifying the relationship between a dependent variable and a combination of several predictors when its underlying regression assumption of normality, homocesdasticity, independence and linearity are satisfied. Failure to satisfy some of this assumption due to high pair wise correlation matrix is the area of concern for this paper. Failure to satisfy this assumption does not mean that our regression model is bad .It means that our solution may under-report the strength of the relationships

The first problem we will discuss is multicollinearity, a tool for model violations in regression assumptions. Multicollinearity occurs when two or more predictors (combination of predictors) are highly (but not perfectly) correlated with each-other. If the variables are significantly alike, it becomes impossible to determine which of the variables accounts for variance in the dependent variable(Y). Hawkings, (1983).

Introduction to regression diagnostics at the level of Chatterjee and Price (1991) or Fox (1991) offer the researcher a variety of numerical and graphical diagnostics for judging the adequacy and violation of a regression model, which include outliers, multicollinearity, nonlinearity, non-independence, heteroscesdasticity and clustering of data points. With the problem of multicollinearity, it is difficult to come up with reliable estimates of their individual regression coefficients, because they basically convey the same phenomenon or information. We have perfect multicollinearity when the correlation between two predictors is equal to 1 or -1. In practice, we rarely face perfect multicollinearity in a real life data set.

We can explore one issue caused by multicollinearity by examining the process of attempting to obtain estimates for the parameter of the multiple regression equation Yi = $\beta o + \beta_1X_{i1} + \beta_2X_{2i} + \beta_3X_{3i} + \ldots\ldots\ldots. + \beta_kX_{ki}$ + ei the ordinary least square estimates involve inverting the matrix X^TX where

$$X = \begin{pmatrix} 1 & X_{11} & . & . & . & X_{k1} \\ 1 & X_{21} & . & . & . & X_{K2} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & X_{kN} & . & . & . & X_{kM} \end{pmatrix}$$

If there is an exact linear relationship (perfect multicollinearity) among the predictors, the rank of X (and therefore of X^TX) is less than K+1 and has determinant 0 and the matrix X^TX will not be invertible. In fact, the X matrix is not of full rank. In most applications, an analyst is more likely to face a high degree of multicollinearity. For example, suppose that instead of the above equation holding, we have that equation in modified form with an error term ei $\beta o + \beta_1X_{i1} + \beta_2X_{2i} + \beta_3X_{3i} + \ldots\ldots\ldots. + \beta_kX_{ki}$ + ei = 0

In this case, there is no exact linear relationship among the variables, but the X_{ij} variables are nearly perfectly multicollinear if the variance of e_i is small for some set of values for the β's. In this case, the matrix X^TX has an inverse, but is ill conditioned so that a given computer algorithm may or may not be able to compute an approximate inverse and if it does so the resulting computed inverse may be highly sensitive to slight variation, in the data (due to magnified effects of rounding error) and so may be very inaccurate

The VIF measures how much multicollinearity has increased the variance of a slope estimate. Suppose that we write the full rank regression model for n independent observations.

$Y_i = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \ldots + \beta_kX_k + e_i$, i = 1… n
where

Tele:
E-mail addresses: reubitosmek@yahoo.com

Var $(e_i) = \sigma^2$. The general linear model is given as Y = βX + e. In vector and matrix approach we have

$$X = \begin{Bmatrix} 1 & X_{11} & . & . & X_{p1} \\ 1 & X_{12} & . & . & X_{p2} \\ 1 & X_{13} & . & . & X_{p3} \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & X_{1n} & . & . & X_{pn} \end{Bmatrix} \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_n \end{pmatrix} \quad Y = \begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix} \quad e = \begin{pmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{pmatrix}$$

The vector Y is the nx1 observation vector, x matrix is called the n x (p+1) design matrix, β is called (p+1) x 1 unknown parameter and e is the nx1 error vector. The name of this diagnostic arises from writings the variance of the least squares estimator β^ (j = 1… k) as (e.g., Belsley 1991, sec. 2.3) Var $(\hat{\beta}_j) = \sigma^2(X'X)\, jj^{-1} = \sigma^2/_{SSj}VIF_j$ where $SS_j = \sum_I (X_{ij}-X_j)^2$ and $VIF_j = {}^1/_{1-R^2 j}$. R^2_j is the R^2 statistic from the regression of X_j on the other covariates. Unfortunately, there is no well defined critical value for what is needed to have a large VIF. Some authors, such as Chatterjee and Price (1991), suggest 10 as being large enough to indicate a problem

Identification of multicollinearity in a set of independent vafriables

Indicators that multicollinearity is present in a regression model

1 The t tests for each of the individual slopes are non significant (P>0.05) but the overall F test for testing all of the slopes are simultaneously 0 is significant (P<0.05).This implies insignificant regression coefficients for the affected variables in the multiple regressions, but a rejection of the joint hypothesis that these coefficients are all zero (using an F-test)

2. The two common measures of assessing both pair-wise and multiple variable collinearity are (i) the tolerance and (ii) Its inverse- the variance inflation factor (VIF). The measures tell us the degree to which each predictor is explained by other predictors. In simple terms, each predictor becomes a dependent variable and is regressed against the remaining predictors. Tolerance is the amount of variability of the selected predictors not explained by the other predictor. Some authors have suggested a formal detention – Tolerance value or the variance inflation factor (VIF) for multicollinearity. Tolerance = $1-R^2_j$, VIF = ${}^1/_{\text{tolerance}}$ = ${}^1/_{1-R^2_j}$ where R^2j is the coefficient of determination of a regression of explanatory j on all the other explanatory. Brien, o (2007) proposed a Tolerance value of less than 0.20 or 0.10 and or a VIF of 5 or 10 and above to indicates a multicollinearity problem. A VIF > 10 indicates that the multicollinearity is unduly affecting the least squares estimates of the regression coefficients minimally with the predictors.

3 The simplest and most obvious means of identifying multicollinearity is the proper examination of the pair-wise correlation matrix for the predictor. The presence of high correlation (generally those of 0.90 and above) is the first indication of substantial collinearity.

Consequeces/effects of multicollinearity

1 Multicollinearity increased the standard errors of the regression coefficient (decreased reliability). Increased standard errors in turn means that coefficient for some predictors may be found not to be significant different from zero. The regression coefficients although indeterminate but posses large standard errors which means that the coefficients cannot be estimated with great accuracy (Gujarati,1995).In that case, the test of the hypotheses that the coefficients is equal to zero leads to a failures to reject the null hypotheses.

2 Multicollinearity does not reduce the predictive power or reliability of the model as a whole or at least within the sample data themselves. It only affects calculations regarding individual predictors. However, if your goal in a multiple regression analysis is simply to predict Y from a set of X variables, then multicollinearity is not a problem. The predictors will still be accurate and the overall coefficient of determination $(R^{2)}$ and adjusted R^2 quantifies how well the model predicts Y values. If your goal is to understand how the various X variables impact 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 coefficients will be very wide

Sources of multicollinearity

Mason et al (1975) enumerated four possible sources of multicollinearity

(i) The data collection method employed

(ii)Constraints on the method or in the population

(iii) Model specifications

(iv) An over defined model

It is important to understand the differences among these sources of the multicollinearity, as the recommendations for analysis of the data and interpretation of the resulting model depends to some extent on the cause of the problem .The data collection method can lead to multicollinearity problems, when the analyst samples only a subspace of the region of the predictors defined in the equation

Constraints of the model or in the population being sampled can cause multicollinearity. An over defined model has more predictors variables than number of observations .These models are sometimes encountered in medical and behavioral research, when there may be only a small number of subjects (sample units) available, and information is collected for a large number of predictors of each subject .the usual approach to dealing with the multicollinearity in this context is to eliminate some of the predictor variables from consideration when r^2 is high and VIF is high. Then there is serious multicollinearity problem. With multicollinearity problem, the value X^1X is singular

Remedies of multicollinearity problem

1 When two variables are highly correlated, say a correlation of 0.90 and above, either of each should be removed from the model. The model with lesser R^2, lesser adjusted R^2 and high standard errors should be removed from the model.

Detecting multicollinearity in a multiple regression model with simulated data

Example 1

Let return to the blood pressure data in which researchers observed the following with high blood pressure

Y = Blood Pressure: X_1 = Weight: X_2 = Duration of Hypertension: X_3 = Stress: X_4 = Age: X_5 = Body Surface Area: X_6 = Pulse

Regressing Y on the six predictors variables (x1, x2, x3, x4, x5, x6), we have the following output: R^2 = 0.998. R^2 adjusted = 0.996.Std Error of the Estimate = .40723 .Durbin Watson = 2.249

From Table 1, observed that body surface area (BSA) and Weight are strongly correlated(r = 0.875). Also Pulse appears to exhibit fairly strong marginal correlations with several of the predictors, including Age(r= 0.619), Weight(r = 0.659) and stress(r = 0.566) .On the other hand, none of the pair-wise correlations among age, weight, duration and stress are particularly strong(r<0.40) in each case. Finally, the correlation between BSA, Weight and Pulse with several predictors should be carefully look into

From Table 2, three of the variance Inflation factors (VIF) 8.4, 5.3 and 4.4 are fairly large. The VIF for predictor weight for example tell us that the variance of the estimated coefficients of weight is inflated by 8.4 because weight is highly correlated with at least one of the predictors in the model

For the sake of understanding, lets verify the calculation of the VIF for the predictor weight. Regressing the predictor X_1 = Weight on the remaining five predictors(X_2, X_3, X_4, X_5,X_6).The result for R^2 = 88.1% or in decimal form 0.881.Adjusted R^2 =0.834, std error = 1.84694. Therefore, the estimated coefficients weight is by definition $V_{weight} = 1/_{1-0.881}$ =8.4. Regressing the predictor X_5 = Pulse on the remaining five predictors(X_1, X_2 X_3, X_4, X_6).The result for R^2 = 0.813.Adjusted R^2 =0.735, standard error = 0.07375

One solution to dealing with multicollinearity is to remove some of the violating predictors from the model. Well, since the Predictors Weight and BSA are highly correlated(r=0.875).We choose to remove either predictors from the model. The decision of which one to remove is often a scientific or practical one. For example, if the researches here are interested in using their final model to predict blood pressure of future individuals, their choice should be clear, which of the two measurements, BSA or weights. The analysis shows that BSA should be removed from the model. Reviewing again the above pairwise correlation in table 1, we see that the predictors Pulse also appears to exhibit fairly strong marginal correlations with several of the predictors, including Age(r= 0.619), Weight(r = 0.659) and stress(r = 0.566), the researcher could also considered removing the predictor Pulse from the model because it has high value of correlations among several predictors. Conclusively, BSA and PULSE was finally remove from the regression model and we are left with four good and reliable predictors (age, weight, duration, stress)

Regressing y = BP on the four remaining predictors age, weight, duration and stress, we obtain the following output

Multiple R = 0.997.R^2 =0.993.Adj R^2 = 0.991.Std Error = 0.53204 and Durbin Watson = 1.663, P value = 0.000

The analysis of Table 3 shows that all values for VIF are below 2 and the Tolerance value are all below 1.This is an indication that there is none existing of collinearity among the predictors, hence, the model is good for prediction

The graph of Table 4 we observed that each of the predictor against the dependent variable(y) and the graph the pairwise correlation matrix, have low correlation. This is an indication that no collinearity exits and the improve model is $y = -15.870 + 1.034X_1\beta_1 + 0.040X_2\beta_2 + 0.002X_3\beta_3\ b3 + 0.684X_4\beta_4 + e_i$ is good for prediction. This shows that model validation of regression assumption of linearity, normality, homoscesdasticity and independence of the observation without any outlying effects. In terms of adjusted R^2 value, we do not seem to lose much by dropping the two predictors BSA and PULSE from our model. The Adjusted (r^2) decreased to only 99.0% from the original adjusted (r^2) value of 99.4 and Multiple R = 0.999 .Our P value = 0.000 actually less that 0.05. Durbin Watson = 1.917 which shows no collinearity

Conclusion

This paper explores on the detention and correction of collinearity diagnostics in a multiple regression analysis. Stimulated data's were used to shows collinearity diagnostics and how it was corrected by removing the two predictors from the model due to high correlation influence. The detentions were show through a high VIF, pairwise correlation matrix and couple with a correction scatters plots. The correlation scatter plots shows substantial collinearity

In all this, removing the predictors serve as a means of corrections and this was accompanied with improved as shown in r^2, adjusted r^2, decrease in standard error and a well behave scatter plots.

We therefore say that removing the effects of collinearity among predictors help to increase regression coefficients, create reliability of the model and position the model for further analysis. It facilitates the job of the analysis.

Finally, I therefore conclude that areas of validation of regression assumption using analysis of residuals should be explores by further researchers

Recommendation

In dealing with collinearity among two predictors in multiple regression analysis, there are other measures in addition to the one stated in this paper on the ways of detecting multicollinearity diagnostics and ways of correcting them

1 Condition Number Test: The standard measure off ill condition in measuring in a matrix is the condition index. It will indicate that the version of the matrix is numerically unstable with finite precision number. The condition number with finite precision numbers. The condition number is computed by finding the square root of (maximum eigenvalue divided by the minimum eigenvalue).If the condition numbers is above 30, the regressions is said to be have significant multicollinearity

2 The complete elimination of multicollinearity problem is not possible but the degree of multicollinearity can be reduced by adopting ridge regression, and principal components regression

3 You can also reduce multicollinearity by centering the predictor variables. To do this compute the mean of each independent variable, and then replace each value with the difference between it. It can be useful in overcoming problems arising from rounding and other computational steps if a carefully program is not used

4 Standardized your independent variables. This may help to reduce a flagging of a condition index above 30

5 Looking at correlations only among pairs of predictors, however is limiting if it is possible that the pairwise correlations are small, and yet a linear dependence exists among three or more variables. That's why many regression analysts often rely on using variance inflation factor to help detect multicollinearity

6 Another check for collinearity is the Durbin-Watson test statistic. Normally, its value should lies between 0 and 4. A value close to 2 suggest no correlation

References

Atkinson, A.C. (1985), "Plots, Transformations, and Regression" , Oxford, U.K,: Oxford Publications

Belsley, D.A. (1991), "Conditioning Diagnostics", New York: John Wiley

Chatterjee, S. and Price, B. (1991), "Regression Diagnostics", : New York: John Wiley

Fox, J. (1991), "Regression Diagnostics", Newbury Park : CA :Sage

Gujarati, D.N.(1988), "Basic Econometrices" , New York : McGraw-Hill

Kutner, Nachtsheim, Neter, "Applied Linear Regression Models", 4th edition, McGraw-Hill, Irwin, 2004.

Marquardt, D.W. (1970), ”Generalized Inverse, Ridge Regression, Biased Linear Estimation, and Nonlinear estimation”,: Techno metrics 12/2) 591,605-07

O” Brien, Robert M.2007, ”A Caution Regarding Rules of Thumb for Variance Inflation Factors” ,Quantity and Quality

Table 1

PAIRWISE CORRELATION MATRIX

	BP	WEIGHT(X1)	DUR(X2)	STRESS(X3)	AGE(X4)	BSA(X5)	PULSE(X6)
BP	1.000	.950	.293	.164	.659	.866	.721
WEIGHT	.950	1.000	.201	.034	.407	.875	.659
DUR	.293	.201	1.000	.312	.344	.131	.402
STRESS	.164	.034	.312	1.000	.368	.018	.506
AGE	.659	.407	.344	.368	1.000	.378	.619
BSA	.866	.875	.131	.018	.378	1.000	.465
PULSE	.721	.659	.402	.506	.619	.465	1.000

Table 2

Regression Coefficients	B	Std. Error	Tolerance	VIF
(Constant)	-12.870	2.557		
WEIGHT	0.970	0.063	1.119	8.417
DUR	0.068	0.048	0.808	1.237
STRESS	0.006	0.003	0.545	1.835
AGE	0.703	0.050	0.567	1.763
BSA	3.776	1.580	0.188	5.329
PULSE	-0.084	0.052	0.227	4.414

Table 3

Coefficients

		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
Model		β	Std. Error	Beta	T	P value	Tolerance	VIF
1	(Constant)	-16.090	3.104		-5.184	.000		
	Weight	1.035	.032	.818	32.759	.000	.812	1.231
	Age	.685	.060	.315	11.507	.000	.674	1.483
	Dur	.045	.063	.017	.712	.489	.841	1.189
	Stress	.003	.004	.017	.666	.517	.808	1.238

Table 4

The graph of pairwise correlation matrix

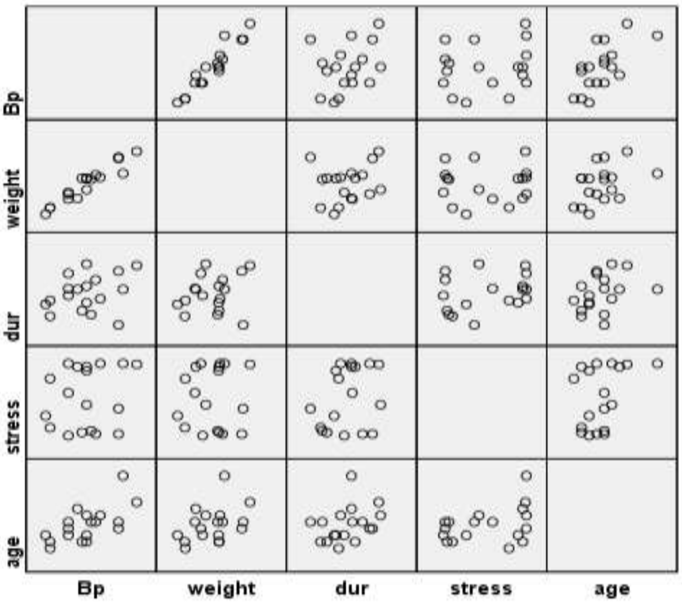

Table 5

Pairwise Correlations Matrix After Removing BSA/Pulse from the model Due to High Collinearity

	WEIGHT	DUR	STRESS	AGE
WEIGHT	1.000	.201	.034	.407
DUR	.201	1.000	.312	.344
STRESS	.034	.312	1.000	.368
AGE	.407	.344	.368	1.000