DEALING WITH VIOLATIONS OF THE ASSUMPTIONS OF CLASSICAL REGRESSION: THE EDUCATOR AND MULTICOLLINEARITY

George Mechling, Western Carolina University

ABSTRACT

The task facing the educator whose charge it is to provide useful instruction in courses which are more quantitative relative to those which for the most part compose business curricula is indeed challenging. Students generally find such courses difficult and unattractive. Consequently, many educators will often regard their instructional efforts successful if at least some of their students grasp a few of the most basic essentials of the topics covered in such courses. Unfortunately however, such "mastery" is not sufficient to adequately support most of the decisions practical decision-makers are called upon to make. One should not therefore, think that even the best of such students are in a position to cut through the complexity and ambiguity of the real world in any sort of analytical way for some useful and practical end. More often than not, what analytical and quantitative tools students have acquired only make them dangerous to themselves and to others who mistakenly rely on them for decision-support. The instruction given to most undergraduate business majors in regression analysis aptly illustrates this dilemma.

Recognizing the practical issues involved in applying regression analyses to real world data, this paper focuses on the phenomenon of multicollinearity. Textbook treatments of multicollinearity tend to be piecemeal in terms of depth and sophistication. Someone interested in studying the subject must range their investigation over a number of sources. This is especially true for those who need to familiarize themselves with the topic but have had little exposure to it. They have little guarantee that such an investigation will be sufficiently complete. To resolve this difficulty, this paper consolidates discussions of multicollinearity drawn from a variety of sources in order to provide readers a reasonably complete and concise coverage of the subject. In this way, educators can have at their disposal a readily available source by which they can make efficient, and hopefully effective, use of their time upgrading the sophistication of their students' regression skills while at the same time, lending meaningful and realistic purpose to their study.

INTRODUCTION

Classical regression requires that a number of assumptions hold if the output such a methodology generates is to provide information upon which one can make decisions. The violation of these assumptions can severely limit this methodology's use. The application of regression analysis to real-world situations all to often involves the violation of one or more

of these assumptions. Textbooks which treat regression analysis will typically attend to at least three of theses violations--autocorrelation, where the error term is not randomly distributed but is serially correlated with itself; heteroscedasticity, where the variance in the error term is not constant across the data set; and multicollinearity, where the variables which make up a regression model's argument are not independent of each other (Gujarati, 1988; Johnston, 1984; Judge, et. al., 1988; Kmenta, 1986; and Myers, 1990).

Most undergraduate business majors receive but one statistics course in their program of studies. Generally, there is barely enough time set aside in such courses to cover simple regression and possibly some multiple regression in which the argument is usually limited to two variables. The student is primarily charged with the rather formidable tasks of conceptually grasping the theories and mechanics of line (or surface) fitting, output interpretation, and reliability testing. Some students may do very well at these tasks. However, the real world is a more complex and ambiguous place than a world for which mastery of these tasks are adequate. Mastery of such tasks certainly prepares students to take on additional tasks whose mastery would better fit them to deal with that complexity and ambiguity. However, the opportunity to do so is often not available to them or it is one they decline to take at the undergraduate level. This is unfortunate. Of the many quantitative tools for conducting analyses of business data to which students are exposed, regression analyses for the purpose of forecasting and estimating the strength of relationships can be one of the most useful. Yet, an analyst unequipped to take into account one or more of the violations of classical regression analysis' assumptions risks not capturing some of the real world's complexity and ambiguity which these violations reflect. Failure to capture this complexity will more often than not generate misinformation and thus, influence decision-making in ways which may lead to undesirable results.

The example data analyzed in this study is real world data and consists of the twenty largest banks in the U.S. as reported by *Fortune* magazine (1985). This data was linearly transformed for instructional purposes and therefore, does not match the data set as it appears in *Fortune*. None of its real-world flavor is, however, lost. The regression output is the same. Only the value of the intercept differs. The data set consists of bank assets, deposits, loans, number of employees, and net income in which this study's interest in multicollinearity begins with how well a bank's net income is explained by the other four variables (Exhibit 1). Since the data is cross-sectional, autocorrelation poses no problem but heteroscedasticity and multicollinearity definitely do. This study will be limited to attending to multicollinearity.

DETECTION OF MULTICOLLINEARITY

Multicollinearity does not exist in simple regression. However, the real world is often sufficiently complex to require more than one variable in the argument of a regression model. When more than one variable makes up the argument, the possibility of multicollinearity becomes a reality. This is particularly true if the analyst loads up the argument of a regression model with variables which have the highest correlation with the dependant variable.

Perfect multicollinearity occurs when at least two regressor variables are perfectly correlated with each other. This yields a data matrix in which two or more columns are exact linear transformations of each other. Thus, the square matrix X'X, is singular and cannot be inverted because the determinant for performing that operation is zero (Gujarati, 1988; Johnston, 1984). Furthermore, such a matrix has infinitely large variances (Gujarati, 1988). This accounts for small t-ratios for parameter estimates of data which is highly self- correlated. The closer to perfect collinearity the data is, the larger the standard errors. Multicollinearity means that the assumption of statistical independence between regressor variables has been sufficiently relaxed to lead to problems of unreliable coefficient estimates whose signs are often reversed. With the relaxation of this assumption, the derivation and application of the least-squares estimator ($\underline{\beta}=(X'X)^{-1}X'\underline{y}$) cannot be justified despite the fact that the least-squares estimator is still unbiased, efficient, and consistent. This is so because multicollinearity is a sample problem and simply selecting another random sample will do little to sidestep the problem. Randomly adding observations to the original sample may help as it increases the degrees of freedom of the estimated model but such a remedy cannot be guaranteed to work nor is it necessarily practical to expand the data set (Kmenta, 1986).

EXHIBIT 1

	Assets ($ Mill.)	Deposits ($ Mill.)	Loans ($ Mill.)	Employees (hundreds)	Net Income ($ Thous.)
Bank #1	287.4500	230.3300	233.6000	196.5600	957.1800
Bank #2	214.7000	175.7700	182.4500	147.2100	820.6800
Bank #3	200.6300	175.5600	165.0000	175.3500	685.2300
Bank #4	151.3700	137.0500	135.1400	138.6000	449.6100
Bank #5	136.8400	127.1100	126.3600	120.7500	330.7500
Bank #6	144.8600	133.2500	128.2100	124.9500	311.0100
Bank #7	125.7300	121.0400	117.5600	119.9100	239.6100
Bank #8	125.2200	121.5900	119.5700	112.5600	354.4800
Bank #9	122.4100	118.1000	116.5300	122.0100	273.4200
Bank #10	121.6100	118.4800	113.4200	121.3800	267.3300
Bank #11	120.0800	117.1000	113.9000	115.7100	241.5000
Bank #12	119.3900	115.1000	113.5900	112.5600	252.8400
Bank #13	117.6600	115.5400	114.0900	112.5600	195.3000
Bank #14	117.4700	114.1800	114.0500	115.9200	225.3300
Bank #15	116.2600	113.1700	111.7800	112.9800	168.2100
Bank #16	115.7500	113.7600	109.0700	108.9900	161.4900
Bank #17	114.7000	113.0200	109.6600	113.1900	170.1000
Bank #18	115.0000	113.3800	110.5400	113.6100	165.2700
Bank #19	113.5700	111.2400	109.8500	113.1900	174.5100
Bank #20	113.3200	111.5900	109.8300	110.6700	173.0400

Upon running a regression on the data using the model:

Bank Profits = f (Loans, Deposits, Employees, Assets)

an examination of the output (Exhibit 2) discloses several troublesome observations:

Exhibit 2

OLS Estimation, 20 Observations, Dependent Variable = Income, Sample Range Set To 1, 20

R-Square = .9468 | R-Square Adjusted = .9326

Variance of the Estimate = 3511.6 | Standard Error of the Estimate = 59.258

Mean of Dependent Variable = 330.84 | Log of the Likelihood Function = -107.140

Analysis of Variance from Mean

	SS	Df	MS	F
Regression	.93716e+06	4	.23429e+06	66.720
Error	52673.	15	3511.6	
Total	.98984e+06	19	52097.	

Variable	Coefficient	Std Error	T-ratio	Correlation	Std Coef	Elasticity
Assets	12.362	11.258	1.0981	.2728	2.4184	5.2199
Deposits	-11.584	12.627	-.91742	-.2305	-1.5350	-4.5454
Loans	-.76948	9.3131	-.82623e-01	-.0213	-.10598	-.29703
Employee	1.9853	3.4033	.58336	.1489	.19999	.75269
Constant	-43.050	535.10	-.80453e-01	-.0208	.00000	-.13012

Variance-Covariance Matrix of Coefficients

	Assets	Deposits	Loans	Employee	Constant
Assets	126.74				
Deposits	-107.41	159.44			
Loans	-79.410	18.208	86.734		
Employee	3.6502	-26.787	12.456	11.582	
Constant	5921.7	-4657.4	-3909.3	-76.101	.28633e+06

Correlation Matrix of Coefficients

	Assets	Deposits	Loans	Employee	Constant
Assets	1.0000				
Deposits	-.75562	1.0000			
Loans	-.75739	.15483	1.0000		
Employee	.95271e-01	-.62335	.39301	1.0000	
Constant	.98300	-.68932	-.78447	-.41789e-01	1.0000

Durbin-Watson = 1.4826 Von Neuman Ratio = 1.5607 Rho = .16740 Residual Sum = -.31051e-11

Residual Variance = 3511.6 Sum of Absolute Errors = 815.33 R-Square Between Observed and Predicted = .9468

First, despite the fact that the R^2 is quite high, none of the t-ratios for the regression coefficients are high enough to reject the null hypotheses at any reasonable level of significance that these coefficients are statistically the same as zero. Second, the signs of the coefficients are not altogether consistent with what one would expect. An inspection of the data concludes that all regression coefficients should be positive because the values the independent variables vary positively with the dependent variable's variation. However, the regression output's estimated coefficients are not consistent with this conclusion. The coefficients for deposits and loans are negative. Finally, an inspection of the correlation matrix of coefficients (Exhibit 2) indicates that the estimated coefficients of some of the independent variables are correlated with each other to a fairly high degree--seventy-five and sixty-two percent.

The first technique, VIF or variance inflation factors, while based on the correlations between pairs of regressor variables, is a more exacting approach than merely inspecting the correlation matrix of coefficients (Johnston, 1984). One first constructs a matrix of correlation coefficients solely between pairs of regressor variables. Doing so yields the following matrix:

	Assets	Loans	Deposits	Employees
Assets	1			
Loans	.9985	1		
Deposits	.9981	.9948	1	
Employees	.9579	.9683	.945	1

One can obtain these correlations in several ways, the simplest being through simple regressions of the regressor variables on each other and taking the square root of the R^2s which are subsequently generated. Inverting the above matrix yields:

Assets	1377.8			
Loans		795.18		
Deposits			462.45	
Employees				33.11

If no linear dependency existed between these four regressor variables whatsoever, they would be said to be *orthogonal* to each other. This is a desirable property in which the resulting inversion of the correlation matrix just exhibited would have "one's" down the main diagonal instead of the values it currently has. Thus, we may interpret those values as the factors by which the variances of the regressor variable coefficients are inflated beyond the ideal case.

There is no hard and fast rule as to how large a factor must be to warrant at least some concern for the extent to which multicollinearity is present. However, if any VIF exceeds a value of ten, it is generally believed that the presence of multicollinearity is extensive enough to warrant action being taken to combat it. Inspection of the VIFs associated with the data analyzed in this case indicates that the values they assume are well beyond a factor of ten and that suggests that probably more than one collinearity of some degree of seriousness exists.

The second technique for detecting the extent of multicollinearity's presence focuses on the *eigenvalues* of the X'X matrix (Johnston, 1984). Eigenvalues are scalars which assume values such that the characteristic matrix is zero. As such, a matrix which is orthogonal will have eigenvalues of one. A matrix which is singular and cannot be inverted will have at least one eigenvalue which is zero. Thus, data matrices which exhibit a good deal of linear dependency between two or more regressor variables will have a least one eigenvalue which approaches zero. An inspection of the principal components analysis found in Exhibit 3 discloses the following four eigenvalues from the X'X matrix of the bank data used in this study--3.9316, 0.66003E-01, 0.18721E-02, and 0.47532E-03. At least the last two eigenvalues are quite small in magnitude and suggest that collinearity probably exists.

Exhibit 3

Principal Components Estimation; 20 Observations; Sample Range Set to 1, 20

PC Assets Deposits Loans Employees / cr ev

principal components on four variables				maximum of four variables retained	
eigenvalues		3.9316	.66003e-01	.18721e-02	.47532e-03
sum = 4.000					
condition numbers		1.000	59.56699	2100.1015	8271.4803
cumulative percentage		.98219	.99941	.99988	1.0000
eigenvectors:	vector 1	-.50297	-.50385	-.50089	-.49221
	vector 2	.27588	.11635	.43859	-.84734
	vector 3	-.17001	-.69235	.67277	.19781
	vector 4	.80125	-.50324	-.32269	.02475

Approaches zero is, however, an issue of relative magnitude. The magnitude of the values found in a matrix could already be small: single digit values or less. The eigenvalues resulting from such a matrix might seem close to zero but not small in magnitude relative to the values found in the matrix. To conclude therefore, that multicollinearity is sufficiently present to warrant concern merely on the basis of a few small eigenvalues could be wrong.

Therefore, calculation of a condition number for the correlation matrix (ϕ) provides a more absolute measure of the extent to which multicollinearity is present. Using the first (the largest) and last (the smallest) eigenvalues from (Exhibit 3) yields:

$$\phi = \text{eigenvalue}_{max}/\text{eigenvalue}_{min}$$
$$\phi = 3.9316/0.00047532 = 8{,}271.48$$

Generally speaking, when the condition number exceeds a value of 1,000, one should be concerned about the extent to which multicollinearity is present and has compromised the reliability of one's regression results (Johnston, 1984). One variant of the eigenvalue approach is less problematic for those who may not care to cope with the cumbersomeness of hand/calculator calculation of eigenvalues or lack the software to do so. Regressing one regressor variable on the remaining regressor variables is consistent with how multicollinearity can be characterized:

$$(\phi_1 x_1 + \phi_2 x_2 + \phi_3 x_3 + \ldots + \phi_k x_k + v = 0), \text{ for example;}$$
$$x_2 = -(\phi_1/\phi_2)x_1 - (\phi_3/\phi_2)x_3 - \ldots - (\phi_k/\phi_2)x_k - v$$

where each of the regressor variables could be regressed on the remaining k-1 variables. High R^2 values in any of these regressions would indicate a high degree of linear dependence existing among the regressor variable. As is readily observable in Exhibit 4, the R^2 values for such regressions as they pertain to the bank data, are quite high suggesting that multicollinearity is extensive (Kmenta, 1986; Wang, 1996).

Exhibit 4

Dependent Variable =Assets	R-Square = .9993; R-Square Adjusted = .9991
Dependent Variable = Deposits	R-Square = .9987; R-Square Adjusted = .9985
Dependent Variable = Loans	R-Square = .9978; R-Square Adjusted = .9974
Dependent Variable = Employee	R-Square = .9698; R-Square Adjusted = .9642

THE PRESENCE OF MULTICOLLINEARITY AS A CAUSE FOR CONCERN

The extent to which multicollinearity might be present in one's data, should be a matter for concern. When the classical regression assumptions hold, an unbiased, efficient (best), and linearly consistent estimator ($\underline{\beta}=(X'X)^{-1}X'\underline{y}$) can be derived (BLUE). Curiously enough however, even when multicollinearity is present, the estimates remain BLUE despite the fact that the standard errors of estimated coefficients are inflated. The most immediate implication of this result is that one cannot be certain whether a variable is significantly associated with the dependent variable. Beyond this problem is a more serious problem when it comes to forecasting with out-of-sample data. The coefficients which are estimated are extremely dependent on the specific data set from which they were generated. Thus, if one wishes to use the model for purposes of prediction even within the resultant model's relevant sample range, some predictions may be quite accurate while others may drastically miss the mark. Hence, the regression coefficients are regarded as *unstable* (Johnston, 1984; Wang, 1996).

Exhibit 5 illustrates this phenomenon. Here, the bank data is rank-ordered according to the magnitude of net income and then partitioned into two subsets consisting of odd and even-numbered observations. Thus, the two subsets are reasonably close replications of each other being almost exactly from the same relevant sample range. The first data subset of odd-numbered observations is then used to estimate a regression model. The coefficients estimated for this model are then used to generate forecasts of net income for the sub-set of even-numbered observations. The model's fit to the out-of-sample data subset (the even-numbered observations) relative to the in-sample data subset (odd-numbered observations) experiences deterioration with the MAPE and MAD increasing from 8.8564 and 17.955 to 22.077 and 84.982, respectively (Exhibit 5). Forecasts made under such a circumstance, therefore, would be ill-advised. The significance of this deterioration will become more apparent when a similar comparison is made after multicollinearity in the data has been effectively addressed.

Exhibit 5: Regression Model Results

Actual and Predicted Bank Income with Prediction Intervals

Where Bank Profits = f(Loans, Deposits, Employees, Assets)

Obs	APE	Actual	Predict	MAD	Obs	APE	Actual	Predict	MAD
1	0.532	957.18	962.29	5.0882	2	46.525	820.68	438.87	381.83
3	0.183	685.23	683.95	1.2536	4	31.268	449.61	309.02	140.58
5	4.983	354.38	336.84	17.664	6	19.182	330.75	267.35	63.454
7	4.816	311.01	296.02	14.978	8	13.867	273.42	235.49	37.913
9	6.951	267.33	248.72	18.580	10	24.878	252.84	180.81	71.992
11	2.072	241.5	246.50	5.004	12	13.046	239.61	270.86	31.258
13	10.149	225.33	202.44	22.865	14	32.854	195.30	259.46	64.164
15	8.271	174.51	160.07	14.433	16	7.198	173.04	185.45	12.453
17	17.848	170.1	200.46	30.359	18	5.743	168.21	177.86	9.659
19	29.838	165.27	214.62	49.323	20	22.612	161.49	198.02	36.518
	8.8564 = MAPE		MAD = 17.955			22.077 = MAPE		MAD = 84.982	

REMEDYING MULTICOLLINEARITY

A number of techniques exist for reducing multicollinearity's adverse impact. Some of the more sophisticated alternatives to ordinary least-squares include ridge regression and principal components regression. These techniques reduce the variance of the estimated regressor coefficients. They are, however, somewhat controversial because the estimates the ridge regression generates are biased and principal components are somewhat artificial in nature. Other less sophisticated alternatives include increasing the sample size, centering and scaling of data, transformation of variables, and the use of *a' priori* information. These techniques have their own unique advantages and disadvantages. This study's bibliography contains discussions of these alternatives.

One of the simplest alternatives is the elimination of a regressor variable or variables which seems or seem to be the most highly correlated with the remaining regressor variables. Exhibit 4 consists of regressions of each regressor variable on the others. As a general rule, the regression generating the highest F statistic identifies the variable with which elimination should begin (Kmenta, 1986; Wang, 1996). This variable is assets. Elimination of assets, however, does little to reduce the apparent presence of multicollinearity. Its symptoms are still present. In fact, one would end up eliminating all but one of the independent variables with the conclusion being drawn that the best decision in this case is to select the simple regression which performs the best. Running regressions with all possible subsets reveals that a regression on the functional form, income = f(Assets), generates the least problematic results (Exhibit 6). The R^2 value barely falls from that in the original model (see Exhibit 1).

Exhibit 6: Analysis of Variance; 20 Observations; Dependent Variable = Income

R-Square = .9434						R-Square Adjusted = .9403
Variance of the Estimate = 3110.4						Standard Error of the Estimate = 55.77
	SS	Df	MS		F	
Regression	.93311e+06	1	.93311e+06		299.997	
Error	55987	18	3110.4			
Total	.98910e+06	19	52058.			
Variable	Coefficient	Std Error	T-ratio	Corr	Std Coef	Elasticity
Assets	4.9629	.28653	17.320	.9713	.97129	2.0937
Constant	-362.17	41.927	-8.6382	-.8976	.00000	-1.0937
Sum of Absolute Errors = 866.79						R-Square Between Observed and Predicted = .9434

As a result of addressing multicollinearity in the data in this way, the excessive sensitivity of estimated coefficients to data specific values not members of the "in-sample" data subset is dramatically reduced. Exhibit 7 illustrates this reduction. The data for this exhibit is rank-ordered and partitioned in identically the same way as was done for Exhibit 5 with the exception that "assets" is the only variable remaining in the argument. The first data subset of odd-numbered observations was then used to estimate a regression model. The coefficients estimated for this model were then used to generate forecasts of net income for the sub-set of even-numbered observations. The model's fit to the out-of-sample data subset (the even-numbered observations) relative to the in-sample data subset (odd-numbered observations) evidences little deterioration with the MAPE and MAD increasing from 13.912 and 40.752 to 16.639 and 48.041, respectively (Exhibit 7). Even though the in-sample MAPE and MAD in Exhibit 7 exceeds the in-sample MAPE and MAD of the multicollinear model in Exhibit 5, the out-of-sample (forecasts) MAPE and MAD in Exhibit 7 are well below the out-of-sample (forecasts) MAPE and MAD in Exhibit 5 by differences of 5.438 and 36.941, respectively.

Exhibit 7: Regression Model Results

Actual and Predicted Bank Income with Prediction Intervals where Bank Profits = f(Assets)

Obs	APE	Actual	Predict	MAD	Obs	APE	Actual	Predict	MAD
1	4.272	957.18	998.09	40.894	2	18.691	820.68	667.32	153.4
3	11.949	685.23	603.35	81.866	4	15.627	449.61	379.34	70.261
5	26.535	354.38	260.44	94.065	6	5.299	330.75	313.27	17.529
7	12.456	311.0	349.74	38.738	8	9.415	273.42	247.66	25.742
9	8.709	267.33	244.02	23.280	10	7.466	252.84	233.93	18.874
11	1.837	241.5	237.06	4.436	12	9.664	239.61	262.75	23.154
13	0.046	225.33	225.20	0.1041	14	15.75	195.30	226.06	30.760
15	18.89	174.51	207.46	32.963	16	19.264	173.04	206.33	33.326
17	24.986	170.1	212.61	42.501	18	30.615	168.21	219.69	51.494
19	29.44	165.27	213.97	48.665	20	34.598	161.49	217.38	55.875
	13.912 = MAPE		MAD =	40.752		16.639 = MAPE		MAD =	48.041

The results of this comparison illustrate an important point. Despite the fact that a multicollinear model's performance explaining in-sample variation may be better than a less multicollinear model's performance explaining the same variation, there is no guarantee that the out-of-sample (forecast) performance of the multicollinear model will also be better. Generally, the odds are that it will be not be as good.

CONCLUSION

The reader should keep in mind that these illustrations of coefficient sensitivity are just that: illustrations. Statistical conclusions are not to be drawn from these illustrations. While it would make no difference as to the regression results however the data was ordered for n=20, it does when partitioning the data set into equal in and out-of-sample subsets for the purpose of generating and examining the accuracy of forecasts. A true experiment to test coefficient sensitivity would entail all possible permutations of the data set prior to its partitioning, running the illustration procedure both ways for each permutation, and then tabulating the results to determine the proportion of permutations in which the regression coefficients were data-specific sensitive in order to assess the risk of ignoring multicollinearity. Such results would not be generalizable beyond the data set analyzed since the chance of selecting a permutation for which coefficient sensitivity was a problem would be a function of the degree to which multicollinearity was a problem for that specific entire set.

This study has consolidated a number of concepts and issues surrounding the topic of multicollinearity. In conclusion, a note of caution needs to be sounded. Heteroscedasticity also afflicts these regression results, including those found in Exhibit 7. It has not gone away by virtue of addressing multicollinearity. Thus, one should not conclude that once one problem has been addressed effectively, decisions based on the ensuing results are now free of other problems.

REFERENCES

Fortune (1985). Vol 111, number 12, June 10, 180-183.

Gujarati, D. (1988). *Basic Econometrics*, 2nd ed., New York: McGraw-Hill.

Johnston, J. (1984). *Econometric Methods*, 3rd ed., New York: McGraw-Hill.

Judge, G., R. Hill, W. Griffiths, H. Lutkepohl & T. Lee (1988). *Introduction to the Theory and Practice of Econometrics*, 2nd ed., New York: John Wiley and Sons.

Kmenta, J. (1986). *Elements of Econometrics*, 2nd ed., New York: Macmillan Publishing Company.

Myers, R. (1990). *Classical and Modern Regression with Applications*, Boston: PWS-Kent Publishing Company.

Wang, G.C.S. (1996). How to Handle Multicollinearity in Regression Modeling, *The Journal of Business Forecasting*, Spring, 23-27.