

Regression: Uses and Abuses

Syed Shahabuddin
Department of Management
Central Michigan University, Mount Pleasant, Michigan
shahals@cmich.edu

ABSTRACT

Regression is a critical tool for establishing cause and effect relationships that could have many implications for business, economics, and society. For accuracy, regression results must be tested for autocorrelation, multicollinearity, and heteroscedasticity. The presence of any of them makes the outcome biased and/or inefficient. Unfortunately, many analysts do not know the rules, ignore the rules, or implement them partially. Consequently, their results are inaccurate. This paper shows how these violations affect results. Many models in which one or all of the violations is present were analyzed. Each model was regressed and results were tested for the presence of the violation, and the equation was then tested for bias and inefficiency.

Key words: regression, multicollinearity, autocorrelation, heteroscedasticity, significance

INTRODUCTION

Many researchers use regression analysis to study relationships between or among variables. According to Ezekiel and Fox (1966), “relations are fundamental stuff out of which all science is built” (p. 65). Many researchers cannot prove or disapprove their hypotheses without establishing some type of cause and effect relationship between or among variables, where one variable may be considered to be affecting one or more other variables. This relationship is described in a functional form indicating that there is an expected relationship between the variables, but it does not specify what that relationship is. In many scientific experiments, relationships can easily be tested and confirmed. However, in many social and economic studies, relationships are not clearly evident and cannot be easily established. Thus, using statistics to establish relationships, if any, and then to confirm this relationship is very critical. A common method for establishing relationships is to determine the average relation between or among variables.

To establish an average relationship, one must select carefully which is (are) the cause(s)—i.e., the independent variable(s)—affecting the dependent variable. Once appropriate cause and effect variables are selected, then relationships must be established. For establishing a relationship, a regression method is commonly used. Regression is based on the concept of “least squares.” To run a regression analysis, one can use the least squares method while making sure that all the required conditions of regression are met. If the conditions are met, the estimated parameters will be the best linear unbiased estimate (BLUE) of the regression parameters. BLUE means that the estimates of linear relationships have the smallest variance. The method is based on a theory that the regression line should have the smallest sum of squared errors from actual observation (data). In addition, if the data have a normal distribution, the estimates will also have a normal distribution. However, for correct results of least squares, one must have a precise measurement of the independent variables and must not have outliers. After estimating the trend line, one must answer the following questions:

1. How close is the relationship?
2. How far is the trend value from the true value of the population from which the sample is drawn?
3. Do the estimates satisfy all the conditions of regression analysis?

The measure of the relationship is the correlation coefficient, which is a measure of the covariability between or among variables. The correlation coefficient (r) measures the extent of the relationship between variables. The correlation coefficient only measures linear relationships. Therefore, if a variable has a strong non-linear relationship, it will not show a high correlation. In case of multiple variables or a single variable, the overall relationship is commonly measured by the Coefficient of Determination – R^2 , which indicates the extent of the effect on the dependent variable(s) by the independent variable(s). A further test to measure how far the true value of the observation is from the trend value (Y) requires determining the extent to which each observation is dispersed

in the overall distribution. This dispersion is measured by sigma (σ_{est}), the standard error of estimate.

The regression should give the best estimates of the parameters. Once estimated, they give residuals, e_i , estimators of ε_i , the error term. The residuals indicate how much of the dependent variable is not explained by the independent variable(s). The sum of squares of the residuals gets smaller by adding additional variables; however, that does not mean that regression results improve. Cornbleet and Gochman (1979) suggested that "for the least-squares model to be valid, these residuals should be random and have a Gaussian distribution with a mean of zero and a standard deviation of one" (p. 434). They also suggest that standard deviation of the residual should be constant at every value of X_i , and repeated measurements of Y would have a standard deviation of $S_{y,x}$. The true regression estimate requires the inclusion of all the appropriate variables. Omitting some relevant variables may introduce a specification bias. When a variable is omitted, a part of the explanation is captured by the remaining variable(s). If the omitted variable is highly correlated with another independent variable(s), the extent of bias of its coefficient will be larger; otherwise, no bias exists.

However, the true regression estimate with irrelevant variables increases the variance of the estimate of all coefficients. That is, even though it does not introduce bias in the regression coefficient, it does reduce the precision of the estimate. A better estimator of relationship is the residual variance, which need not decrease with each additional variable as the degree of freedom increases with each additional variable. A regression equation with smaller residual variance will also have a smaller variance of the error of prediction.

To assess the precision of the estimators of regression, standard errors of the regression coefficient(s) must be calculated. The common measure of precision is the variance of coefficient (CV). The smaller the variance, the higher the precision. Precision also depends on the co-movement of the independent variables, i.e., the less correlated independent variables are with each other, the higher the precision of the regression estimates. To determine the extent of the relationship, a t -test is used to test the significance of each independent variable. The t -test shows whether the dependent and the independent variables are linearly related. That is, if the t -test is significant, one cannot reject the hypothesis that the dependent and independent are linearly related.

A good regression estimate requires that successive error terms, ε_i , are independent of each other and are not correlated with any independent variable. As a result, error terms are assumed to be uncorrelated with the independent variable(s). Failure of this requirement means autocorrelation or serial correlation. Serially-correlated error terms have several consequences (Neter et al., 1983):

- regression coefficients are inefficient;
- mean squared error underestimates the variance of the error terms, as a result, the F may be inflated; and
- the standard deviation of the coefficients may be underestimated.

Autocorrelation is common in much of the economic and business research. To detect autocorrelation, one can use many tests, but the most widely used is Durbin-

Watson (*D*) test, which tests the hypothesis of whether correlation between error terms exists or not. There are many reasons for autocorrelation to exist (Koutsoyiannis, 1973):

- omitted independent variables,
- misspecification of model,
- data that are based on time, and
- random causes.

According to Koutsoyiannis (1973), “the source of autocorrelation has a strong bearing on the solution which must be adopted for the ‘correction’ of the incidences of serial correlation” (p. 197).

A crucial requirement of regression estimates is that independent variables must not be highly correlated with each other. If they are, then they have multicollinearity. In most economic studies, some degree of intercorrelation among independent variables is expected due to the interdependence of many economic variables over time. However, high multicollinearity impairs the accuracy and stability of the estimates. Multicollinearity may be caused by variables that move together over time. Also, by including a variable both as a lag and unlagged will also result in multicollinearity. It is well known that time series may have multicollinearity, but cross-sectional data can have it, too. The effect of multicollinearity on the estimate depends on its severity and the importance of the collinear variable. Greater collinearity increases standard error. Many methods are available to test for multicollinearity. Farrar and Glauber (1967) suggested the chi-square test for detection, the F-test for locating a collinear variable, and *t*-test for finding the pattern of collinearity. Carvalho and Cruze (1996) stated that "A high correlation coefficient indicates multicollinearity. However, when the total number of independent variables is greater than two, this condition become only sufficient, but not necessary, and absence of high correlation between two variables does not indicate absence of multicollinearity" (p. 480). To test for this type of multicollinearity, Marquardt (1970) advised using the variance inflation factor (*VIF*) (p. 600). Neter et al. (1983) state that "if *VIF* is greater than 10, it is possible that the minimum squares regression coefficients associated with such values are highly affected by multicollinearity" (p. 340). Associated with each *VIF* is a tolerance statistic, which can be also used to test multicollinearity. Klein (1962) suggested that if tolerances are less than $(1-R^2)$, multicollinearity exists. Belsley et al. (1980) and Silvey (1969) suggested using eigenvalues. Also, Belsley et al. (1980) stated that "The analysis of the eigenvalues can identify the approximate nature of the linear dependency that exists between the variable" (p. 292). They suggest that an eigenvalue close to zero indicates perfect collinearity, and a small value indicates severe multicollinearity. Klein (1962) recommended using R^2 for significance of multicollinearity, and noted that if *VIF* is greater than $1/(1-R^2)$ or if the tolerance value is less than $(1 - R^2)$, and then multicollinearity exists. Montgomery and Peck (1981) suggested using the condition number for determining multicollinearity. The condition number is the ratio of the highest eigenvalue and the smallest eigenvalue, and the condition index is the square root of the condition number. Belsley (1991) suggested that a condition index within a range of 10 to 30 indicates possible problems of multicollinearity, and a greater value indicates

multicollinearity. Montgomery and Peck (1981) provided the criteria shown in Table 1 for the condition number (CN) to measure multicollinearity.

Table 1: Conditions of Multicollinearity

Condition Number	Existence of Multicollinearity
CN < 100	None
100 < CN < 1,000	Moderate to Strong
1000 < CN	Severe

In regression models, the error term is assumed to be the same for all X's as well as their variances. In other words, if the error terms or variances are constant, the relationship is homoscedastic; otherwise, it is heteroscedastic. Heteroscedasticity can result from an extreme value among independent variables, error of measurement, or misspecification of variables, either dependent, independent, or both. Regardless of the reasons, heteroscedasticity results in biased estimators and biased standard errors. However, unless heteroscedasticity is severe, the estimates can be used without serious distortion.

Many tests are available for detecting heteroscedasticity. The most common are listed below:

- Visual inspection can be done by plotting the residuals against fitted values and detecting uneven behavior in the plot.
- When errors are normally distributed, the White's test works well with the linear relationship between the dependent variable and error variance. White's test requires that squared errors must be regressed with all the independent variables, with the squared of the independent variable and with the cross product of the independent variables, and then nR^2 must be calculated. After calculating the White test, it is compared with chi-square. If the $nR^2 >$ chi-square, then heteroscedasticity exists.
- The Breush-Pagan/Cook-Weisberg test is very similar to the White test except that the researcher selects the variables to be included in the equation. It tests whether error variances are equal. A chi-square test is used to determine heteroscedasticity. The test requires that the chi-square $>$ $RSS/2$ to accept homoscedasticity.
- Goldfeld-Quant (G) test is a bit simpler and tests whether the error variance between subgroups of the data is the same. After calculating the G value (ratio of variance of the second group with variance of the first group), compare it with the F value. If $G >$ F , heteroscedasticity exists; otherwise not. The Goldfeld-Quant test requires that the ratio of the variances of the second group to the variance of first group must be less than the F . However, the division of data into subgroups requires that data must have some unique characteristics allowing differentiation of data, e.g. gender, race, etc.

A regression results must satisfy the aforementioned requirements to produce useful, statistically acceptable results. Unfortunately, many authors use few or none of these tests to support their claims about their results, and many editors and referees accept articles without requiring rigorous testing. The problem has been summarized by Gill (1990), who stated that authors and editors tended "to avoid the hard work of matching appropriate statistical procedures to the specific objectives and characteristics of each project." Gill (1990) also blamed the researchers, reviewers, and editors: "too many researchers blithely ignore the obvious structure inherent in their experiment, and fall into the inferential traps of ambiguity and inefficiency" (p. 191). The situation today seems not to have changed since 1990.

Testing closeness, accuracy, and precision requires statistical analysis. The use of the type of statistical analysis depends on the frequency distribution that determines whether to use parametric or nonparametric statistics (Konings, 1982, p. 371). Classical regression analysis and related tests require data that are normally distributed, but non-parametric tests can be applied to both normal and non-normal data. Most evaluation methods for analyzing data should be subjected to the F-test, t-test, least squares analysis, and calculation of correlation coefficients. However, for statistically-acceptable results, the analysis of the data should be subjected to more comprehensive tests. The purpose of this paper is to show how to analyze data using regression. The paper will show the consequences of accepting results that do not meet each requirement. A variety of simulated data have been created for each condition and results analyzed to show consequences of failing to meet each requirement. The regression requirements are evaluated for small samples as well as large samples. The first data set from a small sample satisfies all the requirements.

Model One

A set of six normally-distributed independent variables was created which were then added to generate the dependent variable. The data consisted of 50 observations. The range of standard deviations among the independent variables is between 1.5 and 7.69 with their means between 19.72 and 70.46.

Significance of the Equation and Multicollinearity

To analyze Model One, SPSS was used to regress the data. To test the significance of the regression coefficient of each variable, a *t*-value was generated. The regression coefficients and their *t* values are listed below (Table 2). Each coefficient has a significant *t* value indicating that each variable is significant and contributes to the explanation of the dependent variable. It also indicates that the standard errors are small, resulting in high, significant *t*-values.

Just because the coefficients of equation have passed the *t*-test, this does not imply that one should accept the results. Further testing is needed. The next test that an equation should be subjected to is the test for collinearity. The suggested tests for collinearity are the Pearson correlation coefficients, the *VIF*, condition index, condition number, and

eigenvalues. A commonly used test for determining multicollinearity is the Pearson correlation coefficient between independent variables. Table 3 shows all the SPSS-generated Pearson correlation coefficients.

Table 2: Regression Coefficients and Variance Inflation Factor (VIF) for Model One

	Unstandardized Coefficients		Collinearity	Statistics
	<i>B</i>	<i>t</i>	Tolerance	<i>VIF</i>
X1	1.046	11.591	.833	1.200
X2	.968	21.219	.841	1.188
X3	1.012	27.150	.911	1.097
X4	.961	37.139	.908	1.101
X5	1.029	44.310	.779	1.284
X6	.998	57.172	.874	1.144

The higher the correlation coefficient (plus or minus closer to one), the more serious is the collinearity. On the other hand, the smaller the correlation coefficient, the less serious is the collinearity. As Table 3 indicates, the largest correlation coefficient number among the independent variables is -.38, indicating no or a very small correlation among the independent variables. If there are more than two independent variables, the correlation coefficient test is necessary but not sufficient for detecting multicollinearity, and additional tests must be applied. The tolerance and VIF tests, and Eigenvalue and condition index or condition number are the other required tests. A VIF greater than 10 indicates multicollinearity and the tolerance less than $(1-R^2)$ indicates multicollinearity. Using R^2 in Table 5 ($1-R^2$) results in .005. All the tolerance values in Table 2 are greater than .005, indicating no multicollinearity. The *VIF* values are less than 10, also indicating no multicollinearity. Additional tests using an Eigenvalue close to or above 1 and a condition index number of less than 100 are applied. Table 4 shows both eigenvalues and condition indices meeting the requirements, thus indicating no multicollinearity.

Since the equation passed the two major tests, i.e., of significance and multicollinearity, the equation should next be tested for its overall significance. The significance tests commonly used are the R^2 and F . In this case, the $R^2 = .995$ and $F = 1,301$ of the analysis are shown in Table 5. Both R^2 and F indicate a "good" equation.

Table 6 (the ANOVA) indicates a very small mean squared error and a high F value. These results also show a significant equation.

Table 3: Pearson Correlations for Model One

	Y	X1	X2	X3	X4	X5	X6
Y	1.000						
X1	.010	1.000					
X2	.061	.237	1.000				
X3	.062	.037	.005	1.000			
X4	.410	-.001	-.215	-.067	1.000		
X5	.476	-.380	-.278	-.060	-.076	1.000	
X6	.735	.013	.024	-.293	.132	.129	1.000

Table 4: Collinearity Diagnostics for Model One

	Eigenvalue	Condition Index
Coeff.	6.950	1.000
X1	.015	21.507
X2	.012	24.140
X3	.011	24.955
X4	.006	34.593
X5	.005	38.373
X6	.001	89.799

Table 5: Additional Tests of Significance for Model One

R^2	F	Durbin-Watson
.995	1,301.152	1.810

Table 6: ANOVA for Model One

	Sum of Squares	df	Mean Squared	F
Regression	6,040.111	6	1,006.685	1,301.152
Error	33.269	43	.774	
Total	6,073.380	49		

Autocorrelation of Model One

Despite the fact the equation passed the three tests, however, the equation must also meet the requirement that the random error terms are uncorrelated. If not, they are autocorrelated or serially correlated. To test for autocorrelation, the Durbin-Watson (D) test is generally used. The result of regression significance in Table 5 indicates $D_{50, 5} = 1.81$, which is greater than the upper limit of the Durbin-Watson value for $D_{50, 4, .01}$ of 1.55. That means there is no autocorrelation.

Heteroscedasticity of Model One

The final test each regression should meet is that the error variances should be homoscedastic (homogeneous). If not, then the regression is heteroscedastic. For determining heteroscedasticity in Model One, the White, Breush-Pagan, and Goldfeld-Quant tests were applied.

- The White test requires that squared errors should be regressed with all independent variables, the square and the cross products of the independent variables, and then nR^2 should be calculated. The analysis indicates that the nR^2 , which is 32, is less than 40.1, the chi-square at 27 degrees of freedom and 95 percent, resulting in the rejection of heteroscedasticity.
- The Breush-Pagan test requires that, to accept homoscedasticity, chi-square must be less than $RSS/2$. The result indicates that chi-square, which is 1.15, is greater than 0/2, the $RSS/2$, again resulting in the rejection of heteroscedasticity.
- The Goldfield-Quant test requires that the ratio of the variances of the second group with the variance of first group must be less than the F . The result indicates the ratio of variances is 1.15, which is less than the F value of 4.20, resulting in the rejection of heteroscedasticity.

Based on all the tests, the analysis proves that the equation meets all the requirements of the regression method.

MODEL TWO

A set of seven normally distributed independent variables was created and then added to generate the dependent variable. The data consisted of 408 observations. The range of the standard deviations of the independent variables is between 2.15 and 7.77 with their means between 20 and 80.

Significance of the Equation and Multicollinearity for Model Two

To analyze Model Two, SPSS was used to regress the data. To test the significance of the regression coefficients of each variable, a t -value was generated. The regression coefficients and their t values are listed below (Table 7). Each coefficient has a

significant t value, indicating that each variable is significant and contributes to the explanation of the dependent variable. It also indicates that the standard errors are small, resulting in high, significant t -values.

Table 7: Regression Coefficients and Variance Inflation Factor (VIF) for Model Two

	Unstandardized Coefficients		Collinearity Statistics	Collinearity Statistics
	<i>B</i>	<i>t</i>	Tolerance	<i>VIF</i>
(Constant)	4.225	3.21		
X1	.998	43.56	.988	1.012
X2	1.01	63.18	.986	1.015
X3	1.01	78.07	.990	1.011
X4	.994	100	.970	1.031
X5	1.01	121	.955	1.048
X6	1.000	139	.991	1.009
X7	.996	153	.971	1.030

Just because the coefficients of the equation passed the t -tests, that does not imply that one should accept the results. The next test an equation should be subjected to is the test for multicollinearity using the Pearson correlation coefficients, the VIF , the tolerance, condition index, condition number, and eigenvalues. Table 7 shows VIF and tolerance values for each independent variable. Each VIF is close to one, indicating no or little multicollinearity. Using the tolerance test, the Model Two has $R^2 = .992$, resulting in $1 - .992 = .008$. Since all of the tolerances are greater than $.008$, no multicollinearity exists. The second test for determining multicollinearity is the Pearson correlation coefficient between independent variables. Table 8 shows all the SPSS-generated Pearson correlation coefficients.

Table 8: Pearson Correlations for Model Two

	Y	X1	X2	X3	X4	X5	X6	X7
Y	1.000							
X1	.128	1.000						
X2	.150	-.074	1.000					
X3	.200	-.002	.051	1.000				
X4	.418	.035	-.004	-.072	1.000			
X5	.502	.037	-.089	-.035	.134	1.000		
X6	.499	.013	-.015	-.016	.078	-.029	1.000	
X7	.559	-.071	-.043	-.058	-.011	.130	.027	1.000

As Table 8 indicates, the largest correlation coefficient number among the independent variables is .134, indicating no or a very small correlation among the independent variables. Since there are more than two independent variables, additional tests for testing multicollinearity must be applied, i.e., the eigenvalues, condition numbers or condition indices. An eigenvalue close to or above 1 indicates no collinearity. Table 9 summarizes eigenvalues and the condition indices; based on the rules discussed previously, multicollinearity is not a serious problem, because the equation passed the multicollinearity test based on the *VIF*, tolerance, Pearson correlation coefficients, eigenvalues and condition indices.

Table 9: Collinearity Diagnostics for Model Two

	Eigenvalue	Condition Index
Coeff.	7.941	1.000
X1	.012	25.878
X2	.011	26.680
X3	.010	28.724
X4	.009	29.313
X5	.008	30.774
X6	.007	33.253
X7	.001	82.286

Because the equation passed the two major tests for the significance and multicollinearity, the equation should then be tested for its overall significance. The significance tests commonly used are the R^2 and F . The $R^2 = .992$ and $F = 12330$ of the analysis are shown in Table 10. Both R^2 and F indicate a "good" equation. Therefore, the tests also worked with a large sample size.

Table 10: Additional Tests of Significance for Model Two

R^2	F	Durbin-Watson
.995	12,330	2.285

The ANOVA, Table 11, shows a small mean squared error and a high F value. Both indicate a significant equation.

Table 11: ANOVA

	Sum of Squares	df	Mean Squared	F
Regression	81,620	7	11,660	12,330
Error	3,726	400	.946	
Total	81,998	407		

Autocorrelation of Model Two

Despite the fact that the equation passed the three tests, it must also meet the requirement that the random error terms should be uncorrelated. If not, they are autocorrelated or serially correlated. To test for autocorrelation, the Durbin-Watson (D) test is generally used. The result of regression in Table 10 indicates $D_{400, 4} = 2.285$, which is greater than the upper limit of the Durbin-Watson value for $D_{400, 4, .01}$ of 1.78, which means there is no autocorrelation.

Heteroscedasticity of Model Two

The final test of Model Two is heteroscedasticity. The White, Breush-Pagan, and Goldfeld-Quant tests were applied.

- The White test requires that the squared errors of the regression should be regressed with all the independent, the square of the independent variables, and the cross product of the independent variables, and the results should be compared with chi-square. If the nR^2 is less than chi-square, then

heteroscedasticity exists; otherwise not. The analysis of the data indicates that, in this case, the nR^2 is 35, less than 49.8, which is the chi-squared at 35 degrees of freedom with 95 percent, thus resulting in rejection of heteroscedasticity.

- The Breush-Pagan test requires that to accept homoscedasticity, the chi-square must be less than $RSS/2$. The result indicates that the chi-square, 51.74, is greater than the $RSS/2$, 0/2, thus, resulting in rejection of heteroscedasticity.
- The Goldfield-Quant test requires that the ratio of the variances of the second group with the first group must be less than the F . The result indicates the ratio of variances is 1.07, which is less than the F value of 1.40, thus resulting in rejection of heteroscedasticity.

Therefore, all the tests of the analysis prove the equation meets all the requirements of regression.

The previous models (small and large sample sizes) show what regression results should be when there are no violations of the regression rules. The next section shows the results of regression where violations exist.

MODEL THREE WITH MULTICOLLINEARITY

The data in the previous model were modified to create multicollinearity. That is, Y was created with some independent variables that were highly correlated. As shown in Table 12, X7 and X8 (.96) show a high correlation, and X1 shows a significance correlation with X8.

Table 12: Coefficient Correlations

	Y	X1	X2	X3	X4	X5	X6	X7	X8
Y	1.000								
X1	.132	1.000							
X2	.147	-.052	1.000						
X3	.203	.001	.045	1.000					
X4	.414	.041	.001	-.067	1.000				
X5	.500	.040	-.090	-.031	.130	1.000			
X6	.497	.012	-.010	-.018	.076	-.030	1.000		
X7	.556	-.072	-.049	-.060	-.019	.129	.025	1.000	
X8	.581	.201	-.066	-.056	-.009	.134	.029	.960	1.000

The t -values in Table 13 decreased for some variables. For example, the t -values for X1 dropped from 43.56 to 1.93 and the values for X7 dropped from 153 to 2.16. The coefficients for both X1 and X7 also dropped, from .998 to .233, and from .998 to .255.

However, all the other t -values also changed but not by much. Other values such as VIF for X1 increased from 1.012 to 17.41 and X7 increased from 1.03 to 213.78. The variable X8, which is highly correlated with X7, has a VIF of 221.71.

The tolerance of X1 dropped from .988 to .057 and of X7 dropped from .971 to .005. The variables which were not correlated showed no or little change in both the VIF and the tolerance values. Since the tolerances of X7 and X8 dropped below .007 ($1-R^2$), they indicate multicollinearity.

The condition indices in Table 14 show that X7 increased from 82.29 to 86.52 and X8 is 754.74; the other indices did not change. The eigenvalues and the condition indices of the other variables did not change much. These measures can easily identify multicollinearity.

Table 13: Coefficients for Model Three

	Unstandardized Coefficients		Collinearity Statistics	
	B	t	Tolerance	VIF
Constant	3.950	2.37		
X1	.233	1.934	.057	17.412
X2	.992	49.10	.982	1.018
X3	1.026	62.14	.989	1.011
X4	.998	78.97	.969	1.032
X5	1.013	96.18	.952	1.051
X6	.999	109.83	.991	1.009
X7	.255	2.16	.005	213.785
X8	.745	6.33	.005	221.452

Table 14: Collinearity Diagnostics for Model Three

	Eigenvalue	Condition Index
Coeff.	8.937	1.000
X1	0.014	25.536
X2	0.012	27.798
X3	0.011	29.075
X4	0.009	30.855
X5	0.009	31.676
X6	0.008	34.432
X7	0.001	86.951
X8	0	753.925

The other statistics (Table 15) indicate that the R -squared decreased (from .995 to .993) with the additional variable X8. The F decreased from 12,330 to 6,720.

Table 15: Additional Tests of Significance for Model Three

R^2	F	Durbin-Watson
.993	6,720	2.116

Autocorrelation of Model Three

The Durbin-Watson value of 2.116 compared with 1.78 indicates no autocorrelation. Table 16 is the ANOVA, which shows MSR , MSE , and the F . All of them decreased compared to the results in Table 11.

Table 16: ANOVA for Model Three

	Sum of Squares	df	Mean Square	F
Regression	81716	8	10,210	6,720
Error	606	399	1.520	
Total	82322	407		

Heteroscedasticity of Model Three

The final test of the model with multicollinearity is heteroscedasticity. The tests used for determining heteroscedasticity have been discussed previously. The Goldfeld-Quandt test indicates $GQ = .69/.50 = 1.38$, which is greater than $F = 1.35$ value and indicates no heteroscedasticity. The White test shows nR^2 is 32.96. The chi-square with 38 degrees of freedom and 95 percent confidence is 54. Since nR^2 is less than chi-square, homoscedasticity is indicated. Therefore, the model shows no heteroscedasticity.

MODEL FOUR WITH AUTOCORRELATION

For Model Four, the data for the Model Two (original model) were modified to add autocorrelated error. A new dependent variable was created that includes autocorrelated error. The Pearson correlation (Table 17) shows no correlation among variables. The t values of all the coefficients (Table 18) decreased by a large number due to higher standard error. The collinearity statistics in Table 18 (tolerance and VIF) did not change at all. The eigenvalue and condition indices (Table 19) did not change. The R^2 , the F , and the Durbin-Watson (Table 19) statistics all decreased. The R^2 decreased from .995 in

Model Two to .683, the F decreased from 12,330 to 123, and Durbin-Watson decreased from 2.285 to .04. Obviously, Durbin-Watson does prove the existence of autocorrelation.

Table 17: Correlation Matrix for Model Four

	Y	X1	X2	X3	X4	X5	X6	X7
Y	1.000							
X1	.081	1.000						
X2	.136	-.054	1.000					
X3	.177	.001	.046	1.000				
X4	.323	.040	-.001	-.067	1.000			
X5	.410	.039	-.092	-.030	.130	1.000		
X6	.417	.013	-.009	-.019	.077	-.030	1.000	
X7	.472	-.071	-.046	-.061	-.017	.130	.024	1.000

Table 18: Coefficients and Other Statistics for Model Four

	<i>B</i>	<i>t</i>	Tolerance	VIF
(Constant)	-25.853	-1.852		
X1	.808	3.311	.988	1.012
X2	1.122	6.610	.986	1.015
X3	1.133	8.171	.990	1.011
X4	.993	9.353	.970	1.031
X5	1.071	12.113	.955	1.048
X6	1.085	14.202	.991	1.009
X7	1.088	15.786	.971	1.030

Table 19: Collinearity Diagnostics for Model Four

	Eigenvalue	Condition Index
Coeff.	7.940	7.941
1	.012	1.000
2	.011	25.980
3	.010	26.700
4	.009	28.703
5	.008	29.334
6	.007	30.625
7	.001	33.256

Table 20: Test of Significance for Model Four

²	F	Durbin-Watson
.683	123.380	.040

The effect of autocorrelation, however, discussed previously that the regression coefficients are inefficient proved to be true. The *t*-values in Table 18 versus Table 7 are lower due to higher standard errors, and the mean squared errors in Table 21 versus Table 6 went up.

Table 21: ANOVA for Model Four

	Sum of Squares	df	Mean Squared	F
Regression	92,892.369	7	13,270.338	123.380
Error	43,022.798	400	107.557	
Total	135,915.167	407		

Only the White test was used to test for heteroscedasticity. The R^2 of the regressing residuals squared of 408 observations was .069. The nR^2 then is 28.15. Comparing it with Chi-Square of 40.10 at 27 degrees of freedom and 95 percent acceptance proves that there is no heteroscedasticity.

MODEL FIVE WITH HETEROSCEDASTICITY

The data for the Model Two without multicollinearity was modified to add heteroscedasticity. That is, errors with variable variances were created and added to the dependent variable. The Pearson correlation (Table 22) shows no correlation among the independent variables. The coefficients of the equation (Table 23 vs. Table 7) increased, and the t values of the coefficients (Table 23) decreased due to higher standard error. The collinearity statistics in Table 23 (Tolerance and VIF) did not change much. The R^2 , the F , and the Durbin-Watson (Table 25) statistics all decreased. The R^2 decreased from .99 in the Model Two to .697, the F decreased from 12,330 to 123, and Durbin-Watson decreased from 2.285 to .785.

Table 22: Pearson Correlations for Model Five

	Y	X1	X2	X3	X4	X5	X6	X7
Y	1.000							
X1	.096	1.000						
X2	.143	-.053	1.000					
X3	.183	.005	.050	1.000				
X4	.328	.041	-.002	-.064	1.000			
X5	.380	.037	-.095	-.032	.128	1.000		
X6	.486	.013	-.007	-.019	.078	-.029	1.000	
X7	.429	-.072	-.045	-.064	-.017	.133	.022	1.000

Table 23: Coefficients and Other Statistics for Model Five

	B	t	Tolerance	VIF
(Constant)	-839.608	-18.656		
X1	2.993	3.805	.988	1.012
X2	3.758	6.853	.985	1.015
X3	3.802	8.512	.989	1.011
X4	3.287	9.595	.971	1.030
X5	3.313	11.611	.954	1.048
X6	4.192	16.986	.991	1.009
X7	3.271	14.694	.970	1.031

Obviously, Durbin-Watson does prove the existence of autocorrelation. The coefficients of the equations (Table 7 vs. Table 23) went up, *t*-values went down, and the mean squared error went up (Table 11 vs. 26). Only the White test of heteroscedasticity was used to test for heteroscedasticity. The R^2 of regressing the residuals squared of 408 observations was .188. The nR^2 then is 76. Comparing it with the chi-square of 43.14 with 35 degrees of freedom and 95 percent acceptance rate proves that heteroscedasticity exists in Model Five.

Table 24: Collinearity Diagnostics for Model Five

	Eigenvalue	Condition Index
Coeff.	7.941	1.000
1	.012	25.953
2	.011	26.651
3	.010	28.708
4	.009	29.421
5	.008	30.617
6	.007	33.265
7	.001	81.956

Table 25: Test of Significance for Model Five

R^2	<i>F</i>	Durbin-Watson
.697	123	.785

Table 26: ANOVA for Model Five

	Sum of Squares	df	Mean Squared	<i>F</i>
Regression	1,031,126	7	14,7303	131
Error	448204.389	400	1,120	
Total	1479331.332	407		

MODEL SIX WITH HETEROSCEDASTICITY, MULTICOLLINEARITY, AUTOCORRELATION

The data for Model Two (original model) were modified to add autocorrelated error, multicollinearity, and heteroscedasticity. A new dependent variable was created that included all these regression violations.

The Pearson correlation (Table 27) shows that X7 and X8 are highly correlated. The t values of the coefficients (Table 28) decreased due to higher standard error. The collinearity statistics in Table 28 (tolerance and VIF) did increase for X7 and X8, i.e., the tolerance values of both variables X7 and X8 went down and VIF values went up. The t -values in Table 28 vs. Table 7 are much lower due to higher standard errors, and the mean squared errors in Table 31 vs. Table 11 went up from .946 to 2,427. The R^2 , the F , and the Durbin-Watson (Table 30) statistics all decreased. The R^2 decreased from .995 in Model Two to .105, the F decreased from 12330 to 5.82, and Durbin-Watson decreased from 2.285 to .364. Obviously, Durbin-Watson does prove the existence of autocorrelation. The effect of autocorrelation, however, discussed previously that the regression coefficients are inefficient proved to be true.

The Eigenvalue of X7 did not change, but the condition indices of both X7 and X8 went up (Table 29 vs. Table 9).

Table 27: Pearson Correlations for Model Six

	Y	X1	X2	X3	X4	X5	X6	X7	X8
Y	1.000								
X1	.042	1.000							
X2	.064	-.054	1.000						
X3	.032	.001	.044	1.000					
X4	.119	.040	-.003	-.071	1.000				
X5	.117	.040	-.095	-.035	.125	1.000			
X6	.167	.013	-.009	-.020	.075	-.032	1.000		
X7	.197	-.072	-.047	-.063	-.019	.128	.023	1.000	
X8	.199	.203	-.064	-.059	-.010	.133	.027	.960	1.000

Table 28: Coefficients and Other Statistics for Model Six

	B	t	Tolerance	VIF
(Constant)	114.45	1.72		
X1	6.34	1.32	.057	17.549
X2	1.09	1.35	.982	1.018
X3	.815	1.23	.988	1.012
X4	.964	1.91	.971	1.030
X5	.868	2.05	.953	1.050
X6	1.16	3.20	.991	1.009
X7	6.40	1.35	.005	214.597
X8	-4.95	-1.05	.004	222.455

Table 29: Collinearity Diagnostics for Model Six

	Eigenvalue	Condition Index
Coeff.	8.937	1.000
X1	0.014	25.536
X2	0.012	27.798
X3	0.011	29.075
X4	0.009	30.855
X5	0.009	31.676
X6	0.008	34.432
X7	0.001	86.951
X8	.000	753.925

Table 30: Test of Significance for Model Six

R²	F	Durbin-Watson
.105	5.82	.364

Table 31: ANOVA for Model Six

	Sum of Squares	df	Mean Squared
Regression	113,535	8	14,191
Error	972,576	399	2,437
Total	1,086,111	407	

Only the White test was used to test for heteroscedasticity. The R^2 of regressing the squared residuals of 408 observations was .308. The nR^2 is then 125. Comparing it with chi-square 50 at 36 degrees of freedom and 95 percent acceptance proves that there is heteroscedasticity.

These results indicate that any violation of the regression method results in an unacceptable and/or weak model. The results speak for themselves as to how important it is to check for violations of regression rules and either remove the violation or do not use the equation that gives incorrect results.

CONCLUSION

Regression is used for research in biomedicine, economics, and business. The research results are used to make many critical medical, economic, and business decisions that could have implications on the people, economies, and investments. Therefore, failing to properly use the regression results could have many consequences that should be avoided, such as hurting or killing people, causing economies to fail, or investments to be lost. Sadly, many researchers either ignore the consequences of their erroneous analytical methods or are inadequately trained in the proper use of regression and, thus unwittingly accept results that could have dire consequences for decision makers. This paper illustrates the results of each violation of regression analysis and how each violation affects the results, making the results less significant or downright wrong.

Regression is a useful tool for establishing relationships between or among variables. The results are then used to make critical decisions. This paper has shown the effect of each rule and why each rule is critical to finding the best outcome for the researcher and the people to whom the research may apply. The proper use of regression can have beneficial results for people. However, improper use of regression can have negative or even injurious effects on people. Taking the rules of regression lightly will result in erroneous outcomes and may cause many human, economic, or investment losses. Therefore, a researcher must understand the proper use of the regression and must not ignore the rules. Unfortunately, many researchers don't do that.

REFERENCES

- Carvalho, S. P. D., & Cruz, C. D. (1996). Diagnosis of multicollinearity: Assessment of the condition of correlation matrices used in genetic studies. *Brazilian Journal of Genetics*, 19, (3), 479-84.
- Belsley, D. A., Kuh, E., & Welch, R. E. (1980). *Regression diagnostics: Identifying data and Sources of Collinearity*. New York: John Wiley and Sons.
- Belsley, D. A. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression*, John Wiley & Sons, New York.
- Ezekiel, M., & Fox, K. A. (1966). *Methods of correlation and regression analysis* (3rd ed.). New York: Prentice Hall.
- Farrar, D.E., & Glauber, R. R. (1967). Multicollinearity in regression analysis. *Review of Economics and Statistics*, 49, 92-107.
- Gill, J. L. (1990). Uses and abuses of statistical methods in research in parasitology. *Veterinary Parasitology*, 36, 3-4, 189-209.
- Goldfeld, S. M., & Quandt, R. E. (1965). Some tests for homoscedasticity. *Journal of American Statistics*, 60, 539-47.
- Glejser, H. (1969, Mar.). A new test for heteroscedasticity. *Journal of the American Statistical Association*, 64, 316-23.
- Koutsoyiannis, A. (1973). *Theory of econometrics*. New York: Harper and Row.
- Konings, H. (1982). Use of Deming regression in method comparison studies. *Survey of Immunologic Research*, 1 (4) 371-74.
- Montgomery, D. C., & Peck, E. A. (1981). *Introduction to linear regression analysis*. New York: John Wiley and Sons.
- Neter J., Wasserman, W., & Kutner, M. H., (1983). *Applied linear regression models*. Boston: Irwin.
- Silvey, S. D. (1969). Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society, Series B*, 31, 539-552.
- Klein, L. (1962). *An introduction to econometrics*. New York: Prentice Hall.

Bibliography

- Allison, P.D. (1999). *Multiple regression: A primer*. Thousand Oaks, CA: Pine Forge Press.
- Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47 (5) 1287-94.
- Bhargava, A., Franzini, L., & Narendranathan, W. (1982). Serial correlation and the fixed effects model. *Review of Economic Studies*, 49 (4) 533-49.
- Borra, S., & Ciaccio, A. D. (2002). Improving nonparametric regression methods by bagging and boosting. *Computational Statistics & Data Analysis*, 38 (4), 407-20.
- Browne, M. W. (1975). Predictive validity of a linear regression equation. *British Journal of Mathematical and Statistical Psychology*, 28 (1), 79-87.
- Cornbleet, P. J., & Gochman, N. (1997). Incorrect least squares regression coefficients in method-comparison analysis," *Clinical Chemistry*, 25, 432-38.

- Durbin, J., & Watson, G. S. (1951). Testing for serial correlation in least squares regression, II, *Biometrika*, 38 (1/2), 159-77.
- Durto, J. D., & Pinto, J. C. (2007). New multicollinearity indicators in linear regression models. *International Statistical Review*, 75 (1), 114-21.
- El Dereny, M., & Rashwan, N. I. (2011). Solving multicollinearity problem using Ridge regression models. *International Journal of Contemporary Mathematical Sciences*, 6 (12), 585-600.
- El Mogahzy, Y.E., & Broughton, R. M. (1989). Diagnostic procedures for multicollinearity between HVI cotton fiber properties. *Textile Research Journal*, 59, 440-447.
- Epps, T. W., & Epps, M. L. (1977). The robustness of some standard tests for autocorrelation and heteroskedasticity when both problems are present. *Econometrica*, 45 (3), 745-54.
- Ezekiel, M., & Fox, K. A. (1966). *Methods of Correlation and Regression Analysis* (3rd ed.). New York: Prentice Hall.
- Farebrother, R. W. (1980). The Durbin-Watson test for serial correlation when there is no intercept in the regression. *Econometrica*, 48 (6), 1553-63.
- Farrar, D.E., & Glauber, R. R. (1967). Multicollinearity in regression analysis. *Review of Economics and Statistics*, 49, 92-107.
- Freund, R. J., & Littell, R.C. (2000). *SAS system for regression* (3rd ed.). Cary NC: SAS Institute.
- Glejser, H. (1969, Mar.). A new test for heteroskedasticity. *Journal of the American Statistical Association*, 64, 316-23.
- Greene, W. H. (2000). *Econometric Analysis* (4th ed.). Upper Saddle River, NJ: Prentice-Hall.
- Gurmu, S., Rilstone, P., & Stern, S. (1999). Semiparametric estimation of count regression models. *Journal of Econometrics*, 88 (1), 123-50.
- Kmenta, J. (1971). *Elements of econometrics*. New York: McMillan.
- Haitovsky, Y. (1969). Multicollinearity in regression analysis: Comment. *The Review of Economics and Statistics*, 51 (4), 486-89.
- Helland, I.S. (2001). Some theoretical aspects of partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 58, 97-107.
- Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76 (2), 297-307.
- Kaur, G., Arora, A.J., & Jain, V. K. (2012, Sept.). Multiple linear regression model based on principal component scores to study the relationship between anthropometric variables and BP reactivity to unsupported back in normotensive post-graduate females. *Advances in Environment, Biotechnology and Biomedicine*, pp. 373 – 377. From the Proceedings of the 1st WSEAS Conferences on Energy and Environment Technologies and Equipment; Agricultural Science, Biotechnology, and Food and Animal Science; and Biomedicine and Health Engineering. Tomas Bata University, Zlin, Czech Republic. Retried from <http://www.wseas.us/e-library/conferences/2012/Zlin/ENAGROBIO/ENAGROBIO-60.pdf>

- Kleijnen, J. P. C. (1992). Sensitivity analysis of simulation experiments: Regression analysis and statistical design. *Mathematics and Computers in Simulation*, 34, 297-315.
- Liao, D., & Valliant, R. (2012). Condition indexes and variance decompositions for diagnosing collinearity in linear model analysis of survey data. *Survey Methodology*, 38 (2), 189-202.
- Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54 (3), 217-24.
- Marquart, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12, 591-612.
- Mayer, D.G., & Butler, D.G. (1993). Statistical validation. *Ecological Modelling*, 68 (1-2), 21-32.
- Montgomery, D. C., & Peck, E. A. (1981). *Introduction to linear regression analysis*. New York: John Wiley and Sons.
- Muller, H. G., & Stadtmuller U. (1987). Estimation of heteroscedasticity in regression analysis. *The Annals of Statistics*, 15 (2) 610-25.
- Reed, A.H., Henry, R.J., & Mason, W.B. (1971). Influence of statistical method used on the resulting estimate of normal range. *Clinical Chemistry*, 17 (4), 275-84.
- Rosendorff, B. P., (2004, Mar. 14). Democracy and the supply of transparency. University of Southern California. Retrieved from <http://www.nyu.edu/gsas/dept/politics/faculty/rosendorff/Transparency.pdf>
- Roso, V. M., Schenkel, F.S., Miller, S. P., & Schaeffer, L. R. (2005). Estimation of genetic effects in the presence of multicollinearity in multibreed beef cattle evaluation. *Journal of Animal Science*, 83, 1788-1800.
- Rothermel, G., & Harrold, M. J. (1996). Analyzing regression test selection techniques. *IEEE Transactions on Software Engineering*, 22 (8), 529-51.
- Twomey, P. J., & Kroll, M. H. (2008). How to use linear regression and correlation in quantitative method comparison studies. *Clinical Practice*, 62 (4), 529-38.
- Waldman, D. M. (1983, May). A note on algebraic equivalence of White's test and a variation of the Godfrey/Breusch-Pagan test for heteroscedasticity. *Economics Letters*, 13, 197-200.
- Warga, A. (1989). Experimental design in tests of linear factor model. *Journal of Business and Economic Statistics*, 7 (2), 191-88.
- Westgard, J. O., & Hunt, M. R. (1973). Use and interpretation of common statistical tests in method-comparison studies. *Clinical Chemistry*, 19 (1), 49-57.
- Wissmann, M., Toutenbury, H., & Shalabh (2007). Role of categorical variables in multicollinearity in the linear regression model. Technical Report 8, Department of Statistics, University of Munich. Retrieved from http://epub.ub.uni-muenchen.de/2081/1/report008_statistics.pdf
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48 (4), 817-38.

Syed Shahabuddin is a professor of Management at Central Michigan University, Mount Pleasant, MI, USA. He teaches operations management, supply chain, and management science courses. He received his Ph.D. from the University of Missouri, Columbia, Missouri. He has been at Central Michigan University for 32 years and has published two books, and third book is ready to be published. At Central Michigan University, he has chaired two major departments in the College of Business Administration. He has published more than 60 refereed articles in major journals. He presents papers and chairs sessions at National conferences of DSI and INFORMS. He was a Fulbright scholar in 1991. Before coming to Central, he taught at the University of Notre Dame for five years.