ASYMMETRY AND REGRESSION

Philip Harris Monchar, AT&T Long Lines

Background

This paper is concerned with the assumption of symmetry underlying both correlation and regression. In particular, the paper examines some of the practical effects of violating the assumption of symmetry in the independent variable. It is noteworthy that standard statistical textbooks in sociology, psychology, and education seem to give scant explicit regard to the consequences of violating this assumption. In addition, a computerized literature search of the statistical and social science fields was able to identify very few articles dealing with this issue, and not one of them was dated prior to 1975. The problem addressed in this article was encountered in an analysis of telecommunications data that involved a hypothesized quadratic model. In the course of the analysis an unexpected non-orthogonal relationship ($r_{(X_1-\bar{X}_1),(X_1-\bar{X}_1)^2} > .80$) was found. The models to be demonstrated below will show clearly that the failure to symmetrize my data led to the unexpected non-orthogonality and a problem in interpreting the results of the analysis.

Bradley and Srivastava (1977 and 1979) discussed a similar issue to the one being related here. In their papers, they pointed to the deleterious effects of not centering an X about its mean in correlation and polynomial regression. In particular, they took issue with Marquardt and Snee (1975) whom they quoted as saying, "in a quadratic model centering reduces, and in certain situations completely removes, the correlation between linear and quadratic terms." Bradley and Srivastava then showed that as the values of the independent variable became less symmetrical the correlations between the linear and quadratic terms approached unity, even after centering the independent variable about its mean. Stimson, Carmines, and Geller (1978) agreed with Bradley and Srivastava and added that "the correlation of X with its square, for example, can be made to vary from almost -1 through 0 to almost +1 by the simple expedient of adding or subtracting a constant to X." Thus, multicollinearity and an ill-conditioned matrix may still be present even after steps have been taken that supposedly prevent them from occurring.

The purpose of this paper is to develop the issue raised by Bradley and Srivastava in their two articles and to empirically model in more detail the effects of independent variable asymmetry in correlation and quadratic regression equations with a single independent variable. Specifically, it is my intent to demonstrate that the degree of effect is a function of skewness and kurtosis in the independent variable.

Demonstration

Consider the single independent variable quadratic regression equation of the form, $Y=a+b_1X_1+b_2X_1^2$. Given the highly collinear relationship between X_1 and X_1^2 one would find it difficult to determine and to interpret the two coefficients of slope, b_1 and b_2. In addition, the determination of the true correlation between Y and X_1 and between Y and X_1^2 would be problematic. Thus, we will transform the above regression equation to the algebraically equivalent form which according to many will solve the two problems raised earlier in this paragraph: And, $Y^*=a^*+b_1^*(X_1-\bar{X}_1)+b_2^*(X_1-\bar{X}_1)^2$. For Marquardt and Snee (1975), $(X_1-\bar{X}_1)^2$ is orthorgonal to $(X_1-\bar{X}_1)$ whereas Bradley and Srivastava (1977 and 1979) and Stimson, Carmines, and Geller (1978) claim that the orthogonal relationship only exists under conditions of symmetry in X. I follow the latter viewpoint. In fact, the limiting factor for determining orthogonality between X and its square is not viewed as $\Sigma(X_1-\bar{X})=0$, but rather as $\Sigma SGN(X_1-\bar{X})^2 \sim 0$.

Table 1 summarizes the effect of asymmetry on both the sum of deviations and the signed sum of squared deviations about a mean. Table 2 extends the effects summarized in Table 1 to the correlation between the two deviations, $(X_1-\bar{X})$ and $(X_1-\bar{X})^2$.

Table 1: The Effects of Asymmetry on the Sum of Deviations About a Mean

	Symmetry	Asymmetry
$\Sigma(X_1-\bar{X})$	= 0	= 0
$\Sigma SGN*(X_1-\bar{X})^2$	~ 0	0 > X > 0

*SGN=Signum

Table 2: The Effects of Data Distribution on the Correlation Between $(X_1-\bar{X})$ and $(X_1-\bar{X})^2$

	Symmetry					Asymmetry				
X_{1i}	$<<\bar{X}$	$<\bar{X}$	$=\bar{X}$	$>\bar{X}$	$>>\bar{X}$	$<\bar{X}$	$=\bar{X}$	$>\bar{X}$	$>>\bar{X}$	
$(X_{1i}-\bar{X}_1)$	<<0	<0	=0	>0	>>0	<0	=0	>0	>>0	
$(X_{1i}-\bar{X}_1)^2$	<<0	<0	=0	>0	>>0	>0	=0	>0	>>0	
$r_{(X_1-\bar{X}),(X_1-\bar{X})^2}$			= 0				0 > r > 0			

A Monte Carlo sampling experiment was used to create 297 sets of 100 random variates of assorted, yet known, distributional characteristics. Each set of 100 random variates was generated by means of a standard algorithm (Procedure UNIFORM of SAS-79.3). This procedure produced 100 pseudo-random numbers (U_i) from a uniform distribution

over the range zero to one. Each U_i in each set was converted to a simulated Weibull value using the function: $X_i=-\log(U_i)^{1/a}$, where a=0.4 and 0.5 to 5.0 by increments of 0.5. The values, $(X_{1i}-\bar{X}_1)$ and $(X_{1i}-\bar{X}_1)^2$ were generated for each X_{1i} within its data set for each of the 297 data sets. The skewness and kurtosis of each of the 297 Weibull distributions were recorded.

Two new variates, Y_L (linear) and Y_Q (quadratic) were created in each of the 297 data sets according to the algorithms: $Y_L=(X_{1i}-\bar{X}_1)+e_L$; and $Y_Q=(X_{1i}-\bar{X}_1)^2+e_Q$, where e_L and e_Q are random normal variates of equivalent range and standard deviations as $(X_{1i}-\bar{X}_1)$ and $(X_{1i}-\bar{X}_1)^2$ respectively. The correlations between Y_L and $(X_{1i}-\bar{X}_1)$ and between Y_Q and $(X_{1i}-\bar{X}_1)^2$ are approximately 0.7. Next, for each of the 297 Weibull distributions of 100 variates the mean and standard deviation of $(X_{1i}-\bar{X}_1)$, $(X_{1i}-\bar{X}_1)^2$, Y_L and Y_Q were set to 50 and 10 respectively. Their intercorrelations were calculated and recorded. Since we are concerned with modelling the ill-effects for correlation analysis of violating the assumption of symmetry in the X-variable we will examine the relationships between skewness and kurtosis in X and: 1) the correlation between $(X_{1i}-\bar{X}_1)$ and $(X_{1i}-\bar{X})^2$; and 2) the correlation between linear Y and $(X_{1i}-\bar{X}_1)^2$.

Finally, the procedure GLM of SAS-79.3 was used to perform multiple regressions on the models: $Y_L=(X_{1i}-\bar{X}_1)+(X_{1i}-\bar{X}_1)^2$, where Y_L has a true linear relationship to $(X_{1i}-\bar{X}_1)$; and $Y_Q=(X_{1i}-\bar{X}_1)+(X_{1i}-\bar{X}_1)^2$, where Y_Q has a true second degree polynomial relationship to $(X_{1i}-\bar{X}_1)$. In order to model the ill-effects for regression analysis of violating the symmetry assumption for the independent variable we chose to examine the relationships between skewness and kurtosis in X and 1) the size of the b coefficient in the equation, $Y_Q=(X_{1i}-\bar{X}_1)+(X_{1i}-\bar{X}_1)^2$; and 2) the change in R^2 due to the second degree polynomial term in the regression on quadratic Y.

Table 3 below summarizes the ranges of values for each of the variables of interest. First, the 297 data sets were ranked in ascending order of their skewness of distribution. Then, the 297 data sets were divided into 9 equal groups of 33 and nine median values for each variable were recorded.

Table 3: Median Values of Skewness, Kurtosis, Correlation Between X and its Square, Correlation Between Y_L and X^2, Standard Error of Slope for X^2 on Y_Q, and change in R^2 for X^2 on Y_Q for 9 Subsets of Data Ranked According to Skewness (N = 297).

	Skew	Kurt	r_{X,X^2} [a]	r_{Y,X^2} [b]	Std. Error [c]	$R^2_{.X^2}$ [d]
	-.36	-.12	-.27	-.18	.075	.02
	-.10	-.39	-.09	-.06	.071	.01
	.13	-.43	.11	.06	.075	.01
	.36	-.22	.26	.18	.074	.03
	.63	.29	.41	.31	.080	.04
	1.08	1.41	.61	.43	.090	.04
	1.87	4.54	.75	.52	.107	.07
	3.46	14.20	.88	.62	.151	.07
	5.55	35.29	.91	.67	.161	.09
Mean	1.43	6.82	.39	.28	.10	.04
S.D.	1.98	14.40	.41	.29	.04	.03
Skew	1.53	2.93	-.28	-.22	1.93	1.32

a: $X=(X_1-\bar{X}_1)$ and $X^2=(X_1-\bar{X}_1)^2$

b: Y = Linear Y

c: Std. Error = s.e. of slope for the regression of $(X_1-\bar{X}_1)^2$ on Y_Q

d: $R^2_{.X^2}$ = Change in R^2 for the regression of $(X_1-\bar{X}_1)^2$ on Y_Q

Table 4 below lists the intercorrelations amongst the above variables. Even a cursory glance at the graphs which follow later clearly shows that most of the relationships, while significant, are not linear. However, as will be demonstrated later the assumption of linearity is quite acceptable. Thus, the correlation coefficients in Table 4 generally understate the strengths of the relationships in the range of values of interest. The following graphs will show the effects of the data distribution of the independent variable on several aspects of correlation and regression. The correlations between $(X_1-\bar{X}_1)$ and $(X_1-\bar{X}_1)^2$, and between Y_L and $(X_1-\bar{X}_1)^2$; the standard error of slope for $(X_1-\bar{X}_1)^2$ on Y_Q; and the change in R^2 due to $(X_1-\bar{X}_1)^2$ being regressed on Y_Q in the full rank quadratic model are illustrated for skewness in Graphs 1a-d and for kurtosis in Graphs 2a-d. The broken horizontal lines in Graphs 1a, 1b, 2a and 2b represent the values of the correlation coefficient ($r \geq \pm .254$) required for the 1% level of significance when Ho: $\rho = 0$ (two-tailed) at the sample size of 100 used in each of the 297 sets.

Table 4: Correlation Matrix for the Variables Skewness, Kurtosis, Correlation Between X and its Square, Correlation Between Y_L and X^2, Standard Error of Slope for X^2 on Y_Q, and the Change in R^2 for X^2 on Y_Q (N = 297).

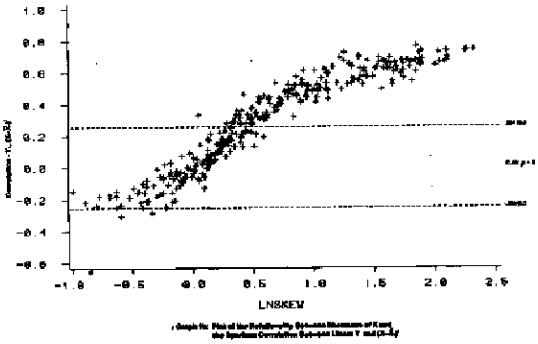
	Skew [a]	Kurt [b]	r_{X,X^2}	r_{Y,X^2}	Std Error [c]	$R^2_{.X^2}$
Skew	1	.90	.96	.94	.88	.58
Kurt		1	.80	.80	.92	.61
r_{X,X^2}			1	.98	.81	.56
r_{Y,X^2}				1	.79	.66
Std Error					1	.48
$R^2_{.X^2}$						1
Mean	.60	1.39	.39	.28	-2.37	.04
S.D.	.75	1.12	.41	.29	.33	.04
Skew	.22	1.06	-.28	-.22	1.16	1.22

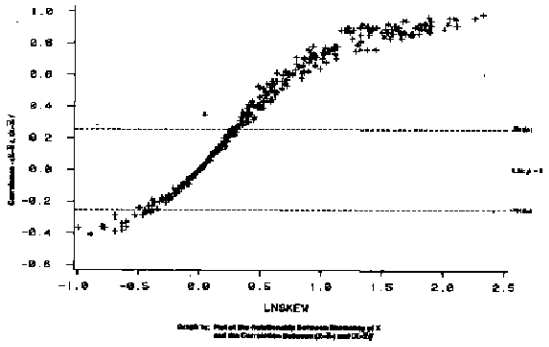
a Transformed value to better symmetrize the data and to reduce heteroscedasticity: Nat. Log(Skewness + 1)

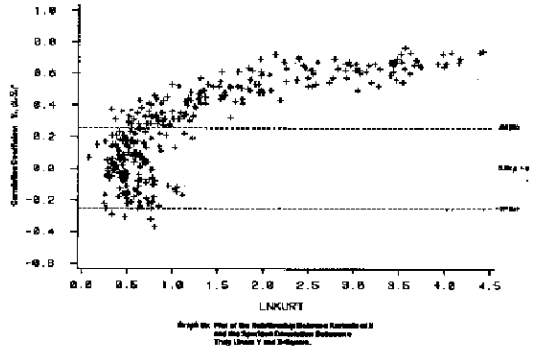
b Transformed value: Nat. Log(Kurtosis + 2)

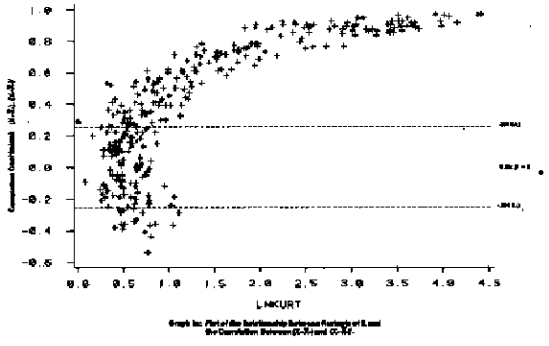
c Transformed value: Nat. Log(Standard Error of Slope)

Graphs 1a and 2a describe the effects of skewness and kurtosis in X on the relationship between $(X_1-\bar{X}_1)$ and $(X_1-\bar{X}_1)^2$ which, according to Marquardt and Snee, is an orthogonal one. Both graphs show that the relationship is not always orthogonal. Whereas the relationship is quite clear-cut for skewness, for kurtosis the relationship is only straightforward at its higher values. For skewness (Graph 1a) quite moderate deviations from symmetry (skew > ± .40) are associated with significant (α = .01) correlations between the first and second-order polynomial terms of X.



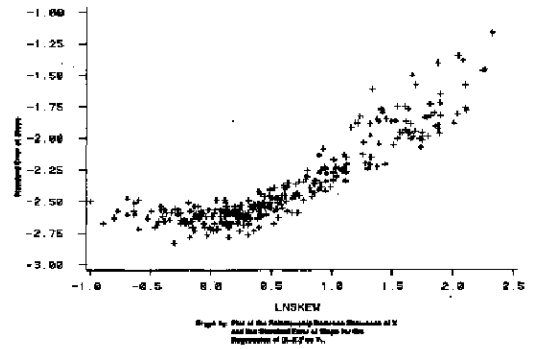


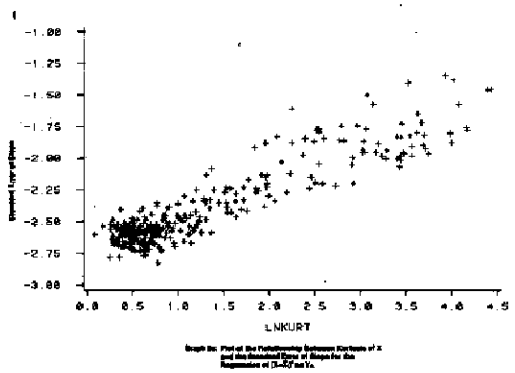




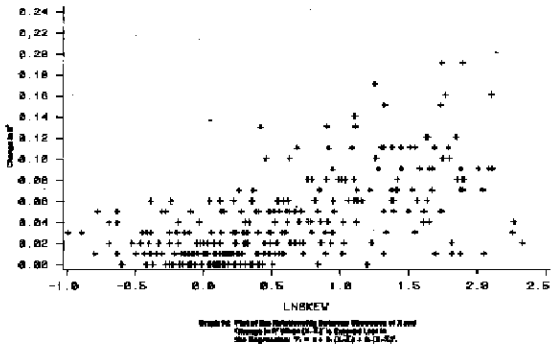
Graphs 1c and 2c illustrate the relationships between the magnitude of the standard error of slope in a regression, in this case a quadratic regression, and skewness and kurtosis in X respectively. For the ranges of skewness and kurtosis present here, the order of magnitude for the standard error of slope is basically unchanging for skewness of -.75 to 1.25 and kurtosis of -1.0 to 1.5. However, for the entire ranges of skewness and kurtosis illustrated in these graphs the magnitude of the standard error of slope changes by a factor of four.

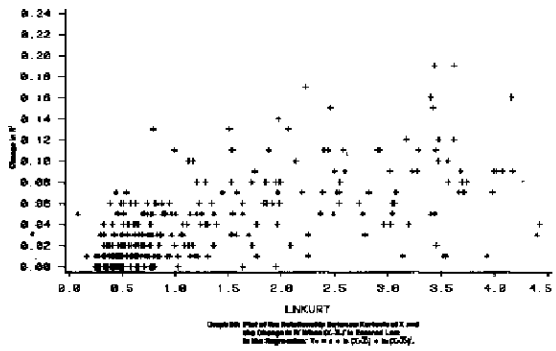
Graphs 1b and 2b show a similar picture for the relationship between the correlation of a truly linear Y to $(X_1-\bar{X}_1)^2$ and skewness and kurtosis in X respectively. Again, the relationships are clearer for skewness than for kurtosis. The significant spurious correlation demonstrated at skewness of greater than -.6 or +.8 is attributed to the relationships: $Y_L=f(X_1-\bar{X}_1)$ and $(X_1-\bar{X}_1)^2=f(X_1-\bar{X}_1)$ when X has an asymmetrical distribution, therefore, $Y_L=f(X_1-\bar{X}_1)^2$.





Finally, Graphs 1d and 2d picture the relationships between the change in R^2 attributable to $(X_1-\bar{X}_1)^2$ in the quadratic equation and skewness and kurtosis in X respectively. These relationships are more moderate than the previous ones, although the relationships remain positive.





Conclusion

I feel that these results emphasize the need to know the distribution of the data. The implicit assumption found in many social science textbooks that only major violations of data symmetry are of concern is shown to be of questionable validity. Rather, the data presented here have indicated that quite moderate deviations from symmetry can induce significant spurious relationships among supposedly orthogonal variables. In general, skewness is sufficient for assessing the dangers of spurious relationships or falsely high standard errors of slope in regression. However, the addition of kurtosis does add a small amount of information which while significant is not really meaningful. I would recommend that the assumption of symmetry be taken seriously and that assymmetrical distributions of variables be transformed such that $\Sigma SGN(X_{1i}-\bar{X}_1)^2 \sim 0$. Symmetrical data distributions have been shown to clearly simplify the correct interpretation of correlation and regression analyses. In conclusion, I feel the effects of data distribution on regression and correlation statistics modelled here are sufficiently important to warrant explicit notes of caution in statistics textbooks.

Bibliography

1. Bradley, R. A. and S. S. Srivastava, "Correlation in Polynomial Regression." Florida State University Technical Report No. M409; Office of Naval Research Technical Report No. 111 (Contract No. N00014-76-C-0394), March 1977.

2. Bradley, R. A. and S. S. Srivastava, "Correlation in Polynomial Regression." The American Statistician, 33, (1): 11-14, February 1979.

3. Cramer, E. M. and M. I. Applebaum, "The Validity of Polynomial Regression in the Random Regression Model." Review of Educational Research, 48, (4): 511-15, Fall 1978.

4. Lawrence, K. D. and D. R. Shier, "A Comparison of Least Squares and Least Absolute Deviation Regression Models for Estimating Weibull Parameters." Communications in Statistics, Simulation and Computation, (Forthcoming 1981).

5. Mosteller, F. and J. W. Tukey, Data Analyses and Regression. Reading, Mass.: Addison-Wesley Publishing Co.: 285-286, 1977.

6. SAS User's Guide. Raleigh, North Carolina: SAS Institute, Inc., 1979.

7. Stimson, J. A., E. G. Carmines and R. A. Zeller, "Interpreting Polynomial Regression." Sociological Methods and Research, 6, (4): 515-24, May 1978.

8. Vasu, E. S., "The Effect of Multicollinearity and the Violation of the Assumption of Normality on the Testing of Hypotheses in Regression." Ph.D. Dissertation (Unpublished), Dept. of Statistics, Southern Illinois University, 1975.