A NOTE REGARDING EVALUATION OF MULTIPLE REGRESSION MODELS

GREGORY N. ALFF

Abstract

Econometric multiple regression models are now commonplace aids to understanding variables affecting the insurance industry. For actuaries and other corporate management personnel to utilize these models to fullest advantage, it is necessary to be familiar with important regression statistics and to be able to critically evaluate model structure.

This paper discusses statistics for determining the strength or validity of a model. Special emphasis is given to the definition of the $\overline{R}^2$ statistic and its relationship to the R^2 and F statistics.

Exclusion of constants from causal models is recommended. Reasons for modeling change in dependent variable rather than level of the variable are considered.

> *"Who reads incessantly, and to his reading brings not*
> *A spirit and judgment equal or superior, . . .*
> *Uncertain and unsettled still remains,*
> *Deep versed in books and shallow in himself."*
>
> John Milton
> *Paradise Regained*

I. THE NEED FOR MODELING

It is not surprising to see rapid growth in the field of econometric research and modeling. Corporate management requires tools to enable it to evaluate economic projections and the probable consequences of alternative marketing and pricing decisions. Work has begun in this area. Econometric models of trends for rate making are now being formulated and utilized for exposures, claim severity, and claim frequency for many lines under the auspices of ISO. Actuaries on industry rate making committees have realized that neither linear nor exponential least squares procedures can be totally relied upon to yield realistic estimates of future trends in today's economic environment. What is needed is an understanding of the causal relationships between outside economic elements and those elements important to insurance rate making and pricing. One vehicle that can provide this understanding is the multiple regression model. In order to make more effective use of the models being developed, it is necessary to be familiar with important regression statistics and to be able to critically evaluate model structure.

II. TOOLS FOR EVALUATION

Actuaries and all levels of insurance management are continually being presented with new, purportedly improved, and ever-more complicated models. In their paper [1] Lommele and Sturgis discuss seven tests for determining the strength or validity of a model. They are as follows:

1. A t-test at the 95 percent level is used to test the importance of each independent variable. The usual standard for this test is $|t| \geq 2$ given at least 16 observations.
2. The sign of the t-test, indicating whether the independent variable's relation to the dependent variable is direct or inverse, should make good intuitive sense.
3. R^2, the coefficient of multiple determination, is a measure of the part of the variation in the dependent variable that is explained by the variation of the independent variables. There is no generally accepted standard of quality for R^2, rather it provides a measure for comparison of one model against another [2]. However, subjective standards do exist and are discussed in the next section in terms of $\overline{R}^2$.
4. The Durbin-Watson d statistic is used to test for autocorrelation in the residual or error terms. The d statistic is generally considered acceptable if $1.5<d<2.5$. A d outside this range would indicate probable serious autocorrelation of error terms.
5. Mean absolute error is an indicator of historical and recent accuracy. A more commonly calculated value is what is often referred to as the standard error of regression. It is calculated as: $SE_R = \sqrt{\Sigma(Y_i - \hat{Y}_i)^2/(N - K)}$, where $(N - K)$ is the degrees of freedom. This is a statistic useful for comparison of models, without a specific threshold for acceptance.
6. Correlation coefficients between each possible pair of variables from a model should show each independent variable to be more highly correlated with the dependent variable than with any other independent variable. If this is not the case colinearity may result, leading to low t-test

values for the two strongly-correlated independent variables as they compete for acceptance in the model.

7. The model as a whole should be intuitively sensible. This test is very important if the model is to gain acceptance with other potential users.

Information for the first five of these tests is often part of the model results presented by computer regression programs and in the published work of econometricians.

All seven tests are important considerations, but even with satisfactory indications from these tests, the model may still contain significant weaknesses.

III. R-BAR-SQUARED ($\overline{R}^2$)

If R^2 for a given model is .93, the person evaluating the model may be very impressed with the model. However, it is possible that he is being deceived. A better measure of fit is $\overline{R}^2$, which is R^2 adjusted for degrees of freedom [3]. Using $\overline{R}^2$ instead of R^2 guards against a model being "overspecified." Being "overspecified" basically means that the model has too many independent variables in conjunction with the given number of data observations, creating a problem with regard to degrees of freedom. A hint of this may come from the t-tests. If the t-test shows a marginal value or a value lower than acceptable at the 95 percent confidence level for a variable, overspecification may be the reason. Sometimes extra variables with questionable t-tests are left in the model because they improve the R^2. The $\overline{R}^2$ statistic will aid in evaluation of whether all variables should be allowed to remain in the model. Extra independent variables will often increase R^2, but $\overline{R}^2$ may decrease if the additional variable has little value.

The reason that $\overline{R}^2$ reacts differently than R^2 is that it is adjusted to account for degrees of freedom. A textbook [4] formula is:

$$\overline{R}^2 = \frac{1 - K}{N - K} + \frac{R^2(N - 1)}{N - K}$$

where: R^2 is the coefficient of multiple determination;
K is the number of independent variables, including any constant;
N is the number of observations.

But algebraically:

$$\bar{R}^2 = \frac{1 - K}{N - K} + \frac{R^2(N - 1)}{N - K} = \frac{1 - K + R^2N - R^2}{N - K}$$

$$= \frac{R^2N - R^2K + R^2K - K + 1 - R^2}{N - K}$$

$$= R^2 + \frac{K(R^2 - 1) - (R^2 - 1)}{N - K}$$

$$= R^2 - \frac{(K - 1)(1 - R^2)}{N - K}$$

Thus, $\bar{R}^2$ is equal to R^2, less a correction for degrees of freedom. Since each of the terms contained in the correction is positive, $\bar{R}^2$ will be less than R^2. The only exceptions are in the special cases when $R^2 = 1.0$ or $K = 1$, where the correction goes to zero and $\bar{R}^2 = R^2$.

The effect on $\bar{R}^2$ and acceptability of t-test values together should determine whether an additional variable is allowed in a model.

There are no generally accepted objective standards of quality for $\bar{R}^2$. However, subjective standards do exist among knowledgeable evaluaters. Such standards vary depending on the variable being modeled and the form and complexity of the model. Prior to examining the details of a simple model for the level of an inflation-sensitive dependent variable, my a priori expectation is that $\bar{R}^2$ should be greater than .90 for the model to be worth reviewing. This is because high values of $\bar{R}^2$ are relatively easy to achieve when modeling the level of such a dependent variable. For a model of change in the dependent variable incorporating a number of complex variable relationships, my expectations of $\bar{R}^2$ will not be as high. For some models of change in the dependent variable, any $\bar{R}^2$ greater than .80 may indicate a model well worth investigating in further detail.

The $\bar{R}^2$ statistic is most meaningful when used as a tool for comparison of competing models. Although $\bar{R}^2$ is an important statistic, it cannot stand alone. All the tests discussed in Section II are important in the evaluation of a given model or when comparing it to alternative models.

IV. $\overline{R}^2$ AND THE F-STATISTIC

Further algebraic substitution into the equation leads to an interesting relationship. The F statistic is defined as:

$$F = \frac{\Sigma(\hat{Y}_i - \overline{Y})^2/(K - 1)}{\Sigma(Y_i - \hat{Y}_i)^2/(N - K)} = \frac{\text{explained variance}}{\text{unexplained variance}}$$

and

$$R^2 = \frac{\Sigma(\hat{Y}_i - \overline{Y})^2}{\Sigma(Y_i - \overline{Y})^2} = \frac{\text{explained variation}}{\text{total variation}}$$

where:

Y_i is the dependent variable for point or year i;
$\hat{Y}_i$ is the fitted value;
$\overline{Y}$ is the mean of the Y_i values.

Then it can be shown that:

$$F = \frac{R^2/(K - 1)}{(1 - R^2)/(N - K)} \qquad \text{(see Appendix I)}$$

and by manipulating this formula,

$$\frac{(K - 1)(1 - R^2)}{(N - K)} = \frac{R^2}{F}$$

so finally,

$$\overline{R}^2 = R^2 - \frac{R^2}{F}$$

or $\overline{R}^2 = R^2(1 - 1/F)$

The F statistic is used to indicate the significance of the entire regression. With 11 or more observations, an $F \geq 5$ indicates a "significant" regression [5]. Note that given $F = 10$, then $\overline{R}^2 = .9R^2$. The example introduced in Section III where R^2 was .93 would be .9 × .93 = .84 when adjusted for degrees of freedom. Such a model may not be quite so impressive when compared to another model that may be better specified by a different set of independent variables, and thus have a higher $\overline{R}^2$.

V. CONSTANT WEAKNESS

It is common in causal models to include a constant term. It is not unusual for the constant to have a strong *t*-test, indicating it is a strong contributor in the explanation of the level of, or change in, the dependent variable. Such a constant often may only be serving as a proxy for an economic variable that has historically shown stability or consistent period-to-period movement (depending on the form of the model equation). In an earlier paper [6] presenting a model of general liability written premium, it was noted that a constant did not improve that model. Rather, the major effect of inserting a constant was to replace one of the independent variables, as indicated by *t*-tests.

A constant does nothing to describe the underlying contributory causes of change in the dependent variable. Any independent variable which seems to have a logical causal effect on the dependent variable should be carefully tested. If the dependent variable and the constant are independently inserted in separate tests of the model, and the *t*-test for the independent variable is similar in strength to that for the constant, then the variable should be preferred. A stronger model may result from the inclusion of an explanatory variable, even if historically stable, because future movements in such a variable may prove important in the usefulness of the model as a predictor.

A constant may be statistically strong, but it does not help "explain" the movement in the dependent variable.

VI. MODELING CHANGE IN THE VARIABLE

Many models being presented use the level of the actual values over time as the dependent variable. In an earlier paper [7], it is suggested that fitting to actual values or levels of an inflation-sensitive variable can often lead to problems such as:

1. Causing colinearity of independent variables;
2. Misestimating turning points; or
3. Masking the true magnitude of error.

It is the third concern which is important in the context of this paper. The following is an example of a least squares linear regression fit to a set of actual values or levels:

Actual Value	
or Level	Fitted Value
200	205.6
220	220.4
245	235.2
260	250.0
250	264.8
275	279.6
300	294.4

$$R^2 = .922$$
$$\bar{R}^2 = .906$$

There is certainly an upward trend and the model appears to produce a good fit. But is management really concerned about the long-term trend, or is it perhaps more concerned with the change from one year to the next? If the concern is with annual changes—how does the above model perform?

	Implied Annual
Annual Change	Changes From
In Actual Values	Fitted Values
+ .100	+ .072
+ .114	+ .067
+ .061	+ .063
− .038	+ .059
+ .100	+ .056
+ .091	+ .053

$$R^2 = .051$$

If the concern is with annual change, there is a need to develop a causal model of annual change that can do a better job of projecting this uneven and possibly cyclical annual change series. This is illustrated by the graphs in Appendix II.

If the purpose of a model is to establish the direction and magnitude of a long-term trend, then modeling with actual value or level as the dependent variable may be sufficient. However, if points of fluctuation, turning points, or the magnitude of any individual points are important, then the model should be based on change in actual values as the dependent variable. In a long-term inflationary environment, modeling level of actual values is relatively easy and high ($>.90$) values of $\overline{R}^2$ should be expected. This is because the magnitude of variable values and underlying long-term trend mask the true annual movement in the dependent variable. As shown in the example above, modeling annual changes instead of level is one approach which will unmask the movement in the dependent variable. Detecting and defining causal relationships for a model of annual change in the dependent variable is more difficult. A model of annual change for a cyclical series in most cases should be preferred to a model of annual level because the value of $\overline{R}^2$ is more meaningful.

Another approach currently being utilized by actuaries working with loss severity trend is the removal of estimated underlying economic trend from the loss severity series by dividing severity values by index values from a deflator such as the GNP deflator. The underlying trend indicated by the indexed deflator is then set aside to be added back later in the analysis. This unmasks the true or residual trend in the insurance loss cost after stripping away the effects of general economic inflation. It is often difficult to develop a causal model with a high $\overline{R}^2$ to fit the residual annual change series. However, a clearer understanding of the causal effects of the independent variables is gained from the regression statistics of such a model.

The $\overline{R}^2$ statistic becomes more meaningful when it is not exaggerated by the effect of underlying long-term trend or general economic inflation.

VII. MODELS IN A DYNAMIC ENVIRONMENT

Even if a model of annual change does well in explaining a long-term historical cyclical pattern, its ability to predict future change should be carefully analyzed. The model of industry general liability premiums contained in the *Proceedings* [8] is a good example. That model fits 20 years of annual change data well. It predicted the first negative annual changes in written premium for 1980 and 1981, but the predicted return to strong positive premium increases in 1982 and 1983 did not happen. The economic environment changed dramatically, and strong surplus positions and industry competition for cash flow have not allowed premiums to rise. The model did include a variable to measure surplus position, but high investment yields and cash flow patterns were not directly accounted for. Did the model fail then? No, it provided an excellent explanation of premium changes for years 1962–1981, but this example clearly points out the need for continual adjustment and modification in a changing economic environment. The model must be modified if it is to be useful in the future. Any model should be reviewed regularly to be sure that the relationships on which the model is based continue to hold true.

Modeling can be used effectively to examine and better understand the relationships between elements in a complex and dynamic economy. This note emphasizes the $\overline{R}^2$ statistic as being one statistic and first difference in actual data as being one approach important to evaluating a multiple regression model. An understanding of important regression statistics and techniques for evaluation of model structure will enhance the usefulness of the modeling tool.

REFERENCES

[1] J. A. Lommele and R. W. Sturgis, "An Econometric Model of Workmen's Compensation," *PCAS*, Volume LXI, 1974, p. 170.

[2] P. Kennedy, *A Guide to Econometrics*, first U.S. edition, MIT Press, 1979, p. 25.

[3] Ibid., p. 52.

[4] J. Johnston, *Econometric Methods*, second edition, McGraw-Hill, 1972, p. 130.

[5] S. D. Wheelwright and S. Makridakis, *Forecasting Methods for Management*, second edition, John Wiley and Sons, 1977, p. 116.

[6] G. N. Alff and J. R. Nikstad, "A Model of Industry General Liability Net Written Premiums," *PCAS*, Volume LXIX, 1982, p. 35.

[7] Ibid., p. 31.

[8] Ibid.

APPENDIX I
DEFINITION OF THE *F*-STATISTIC IN TERMS OF R^2

$$F = \frac{\Sigma(\hat{Y}_i - \overline{Y})^2/(K - 1)}{\Sigma(Y_i - \hat{Y}_i)/(N - K)}$$

$$F = \frac{[\Sigma(\hat{Y}_i - \overline{Y})^2/\Sigma(Y_i - \overline{Y})^2]/(K - 1)}{[\Sigma(Y_i - \hat{Y}_i)^2/\Sigma(Y_i - \overline{Y})^2]/(N - K)}$$

We know that total variation = explained variation + unexplained variation,

$$\Sigma(Y_i - \overline{Y})^2 = \Sigma(\hat{Y}_i - \overline{Y})^2 + \Sigma(Y_i - \hat{Y}_i)^2$$

so

$$\Sigma(Y_i - \hat{Y}_i)^2 = \Sigma(Y_i - \overline{Y})^2 - \Sigma(\hat{Y}_i - \overline{Y})^2$$

and

$$R^2 = \frac{\Sigma(\hat{Y}_i - \overline{Y})^2}{\Sigma(Y_i - \overline{Y})^2}$$

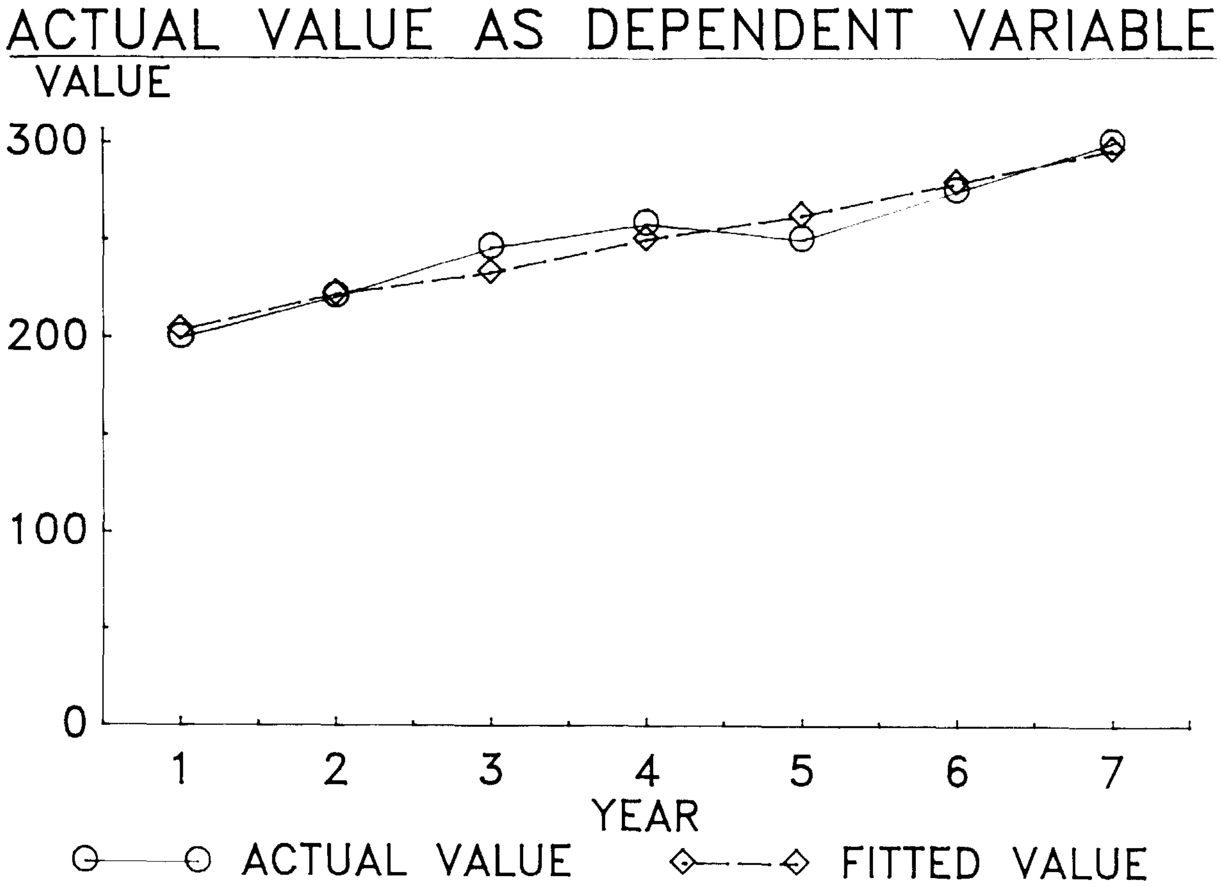
then by substitution,

$$F = \frac{R^2/(K - 1)}{\{[\Sigma(Y_i - \overline{Y})^2 - \Sigma(\hat{Y}_i - \overline{Y})^2]/\Sigma(Y_i - \overline{Y})^2\}/(N - K)}$$

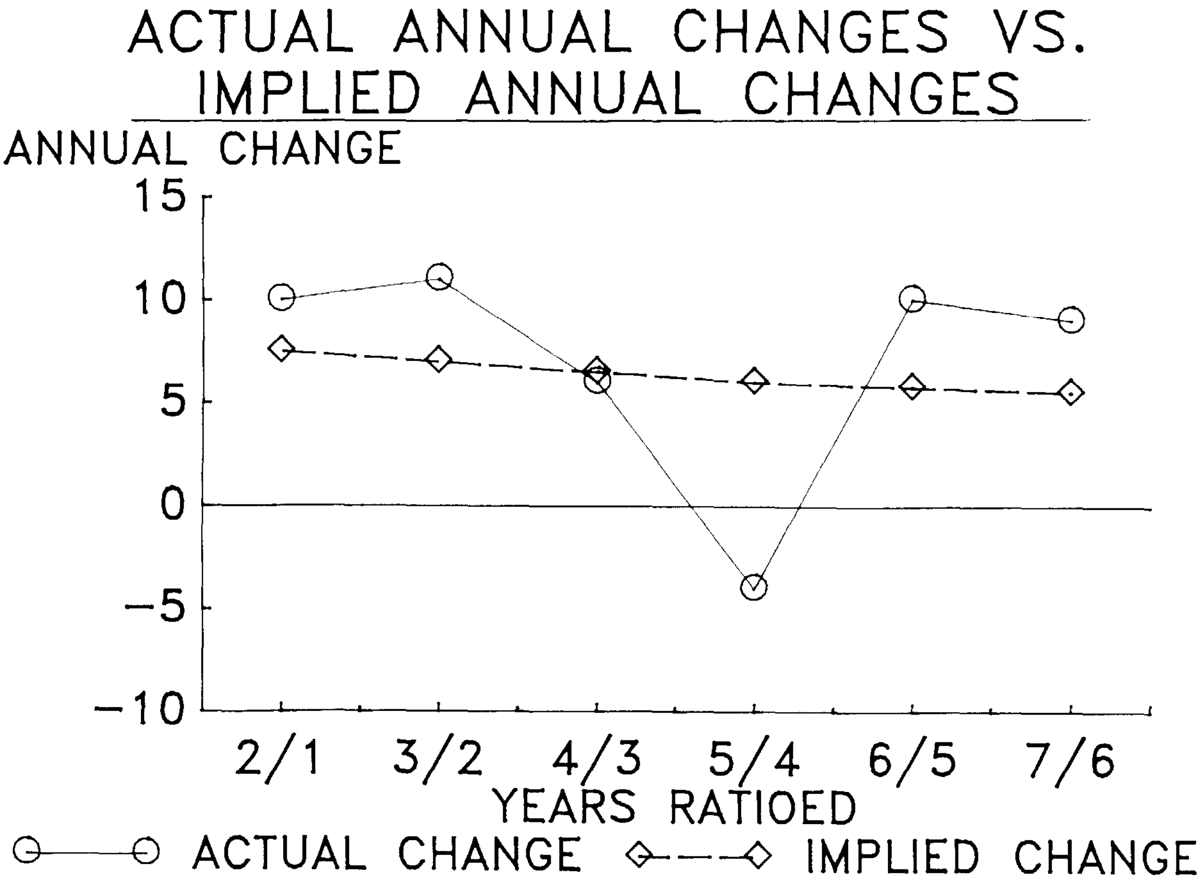
Finally

$$F = \frac{R^2/(K - 1)}{(1 - R^2)/(N - K)}$$

APPENDIX II
GRAPH I



APPENDIX II
GRAPH II


ACTUAL ANNUAL CHANGES VS.
IMPLIED ANNUAL CHANGES
ANNUAL CHANGE
2/1
3/2
4/3
5/4
6/5
7/6
YEARS RATIOED
ACTUAL CHANGE
IMPLIED CHANGE