A STATISTICAL NOTE ON TREND FACTORS: THE MEANING OF "R-SQUARED" (CASUALTY ACTUARIES OF THE NORTHWEST, 3/89)

D. Lee Barclay

A STATISTICAL NOTE ON TREND FACTORS: THE MEANING OF "R-SQUARED"

D. Lee Barclay

Regression models have become standard actuarial tools for analyzing trends in frequency, severity, pure premium, reserves, development factors, and so on. Such analysis often is the basis for estimating future values of these random variables as an important aspect of ratemaking and reserving.

Since the inflationary spiral of the 1970s, the exponential curve has replaced the straight line as the regression model of choice. The exponential model is now commonly accepted even by regulators. By fitting an exponential curve, we actuaries can avoid the underestimation of losses that often results from the decreasing rate of change that is characteristic of the linear regression model. However, linear and other polynomial regression models are still used in some situations. Occasionally, other families of curves, such as logarithmic curves or power curves, are suggested as appropriate models.

In most cases, the purpose of the regression model is to obtain a "trend factor" that accurately reflects what has happened and/or will happen during the time period that interests us. In the linear model, the trend "factor" is a constant amount of increase or decrease per year. When we fit an exponential curve, we look for a constant percentage of annual increase or decrease.

Our models yield several tools that are useful for checking the validity of the trend factor. It is worthwhile to consider the magnitude of the residuals (mean squared error, for example) and whether the residuals show any discernible patterns over time. But the statistic that is used most often is the coefficient of determination, commonly called "R-squared." In imprecise terms, the coefficient of determination is the proportion of the data's variability over time that is explained by the fitted curve. But we often use this statistic as a measure of how well our model fits the data. If the coefficient of determination is high (near one), we are happy and our job is done. If it is low (near zero), we consider the model—or perhaps the data—nearly useless, and we look around for something else that will serve the same purpose.

A quotation from an actuarial software manual illustrates this common view: "This statistic [R-squared] indicates how good the fit of the line or curve is to the data points. A zero R-squared implies a poor fit of the line or curve to the data. . . ." And a large insurer has used the coefficient of determination as the maximum credibility it would assign to a trend factor.

Unfortunately, the coefficient of determination, by itself, is a poor measure of goodness-of-fit.

<u>Low R-Squared/Good Fit</u>

Consider this example, using the linear model for simplicity. Example 1 shows "data" for 10 years. The datum for each year is an independent observation from the normal distribution with mean 50 and variance 1. One would not expect to see a significant trend in these data, and, indeed, the slope of the fitted line is near zero. Although we can see from the residuals that the line fits pretty well, the coefficient of determination is only .024. (Note: Graphs of all examples are appended following the text.)

Example 1

Linear Model
Distribution: Normal(50,1)

Year	Data	Fitted Line	Residual
1979	48.746	49.425	−0.679
1980	49.914	49.461	0.453
1981	49.246	49.498	−0.252
1982	50.297	49.535	0.762
1983	48.455	49.571	−1.116
1984	50.088	49.608	0.480
1985	50.559	49.645	0.914
1986	50.173	49.681	0.492
1987	49.336	49.718	−0.382
1988	49.084	49.755	−0.671

Slope	0.037
Coefficient of determination	0.024
Mean squared error	0.446

In Example 2, we have introduced a positive trend into the same sample by adding one to the second point, two to the third, etc. (Clearly, this is equivalent to taking the first year's datum from Normal(50,1), the second year's from Normal(51,1), and so on. However, we are avoiding the random differences that would result from using data that are independent from those of Example 1.) We would expect the slope of the fitted line in Example 2 to be near one. It is; in fact, it is exactly one plus the slope in Example 1. The coefficient of determination for Example 2 is .952. But as the residuals are identical to those in the first example, we cannot say that this line fits any better.

Example 2

Linear Model
Distribution: Normal(Year-1929,1)

Year	Data	Fitted Line	Residual
1979	48.746	49.425	-0.679
1980	50.914	50.461	0.453
1981	51.246	51.498	-0.252
1982	53.297	52.535	0.762
1983	52.455	53.571	-1.116
1984	55.088	54.608	0.480
1985	56.559	55.645	0.914
1986	57.173	56.681	0.492
1987	57.336	57.718	-0.382
1988	58.084	58.755	-0.671

Slope	1.037
Coefficient of determination	0.952
Mean squared error	0.446

We could analyze these examples in terms of the equations that are found in basic texts on regression techniques, but it may be more helpful to discuss them less precisely. Both examples have the same amount of random error (also known as "white noise"). A curve that fits the data well explains everything but the random error. In both examples, the straight lines do that pretty well, but in the first one, there is little systematic variation ("trend") to be explained. The actuary should be concerned not with the proportion of the data's variation that is explained but with the magnitude of what is left unexplained. (Note that "magnitude" is still a relative term here; we might view the situation differently if the data in our examples began at five instead of at 50.)

Certainly one could construct counterexamples, but the general rule is this: When the fitted line or curve is steep, the coefficient of determination tends to be large; when the fitted line or curve is nearly flat, the coefficient of determination is likely to be small. But this does not imply that the steep line or curve fits the "steep data" any better than the nearly horizontal line or curve fits the "flat data." And, in particular, the low coefficient of determination does not imply that the relatively flat line or curve fits the data poorly.

High R-Squared/Poor Fit

Another example will show that a high coefficient of determination does not necessarily mean that the selected curve fits the data well. During the 1980s, the rate of inflation decreased substantially. For many lines of insurance, severity and pure premium data for these years reflect this decreasing rate. Still, the exponential model, which assumes a constant annual percentage change, prevails in most actuarial trend calculations.

This presents a problem. The exponential curve has a convex shape. But with inflation decreasing, the data points are likely to follow a curve with a concave shape. Example 3 shows the fitting of an exponential curve to "data" that follow a concave power curve. (The "data" are not random here, as the presence of white noise could obscure what is happening.) Even though the exponential curve is the wrong shape, the coefficient of determination is rather high at .946. This fact could easily tempt an actuary to use the exponential curve's trend factor, which is 9.3% per year.

Example 3

Exponential Model
"Data" = 25 + SQRT(Year-1977)

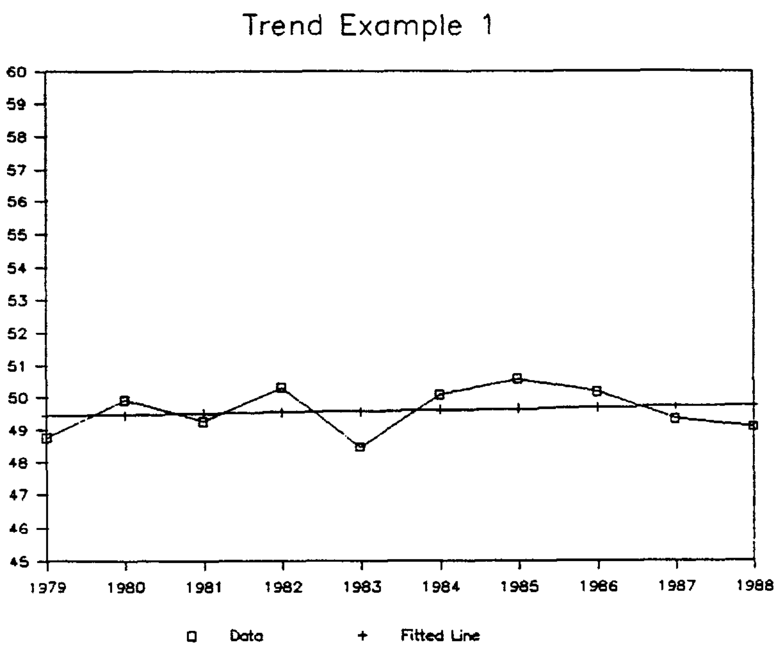
Year	Data	Fitted Curve	Residual
1979	35.355	40.174	-4.819
1980	43.301	43.916	-0.615
1981	50.000	48.007	1.993
1982	55.902	52.478	3.423
1983	61.237	57.367	3.871
1984	66.144	62.710	3.434
1985	70.711	68.551	2.160
1986	75.000	74.936	0.064
1987	79.057	81.916	-2.859
1988	82.916	89.546	-6.631

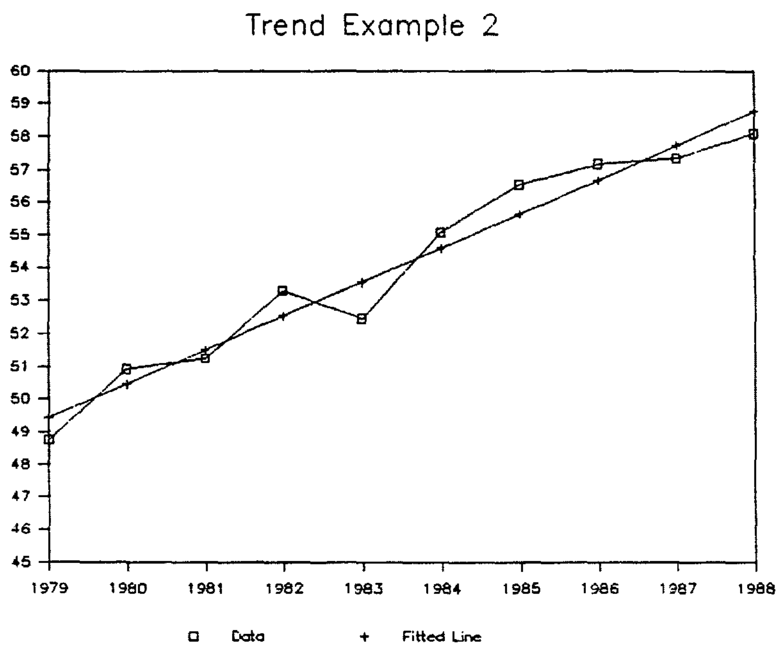
Slope percentage	9.314
Coefficient of determination	0.946
Mean squared error	12.287

The potential for overestimation is significant. If, for instance, these data--for 1979 through 1988--were used in ratemaking, the trend problem might involve making an estimate for 1990. The fitted exponential curve hits 107.0 in 1990, whereas the power curve is at 90.1. Use of the 1990 value from the fitted curve would result in an error of 18.7%. But nowhere between 1979 and 1988 is the difference between curves so large.

Without drawing a graph, one can often detect a poorly fitting curve by looking at the signs of the residuals. In this example, the residuals are negative, then positive, then negative again, following a clear pattern. When a curve fits well, the signs of the residuals will appear to be distributed more randomly.

One obvious solution is to use a more appropriate model--that is, another type of curve. But industry ratemakers--in both companies and rating bureaus--tend to use the exponential model regardless of how poorly it fits. The underwriters and marketers may then adjust the actuarial indications downward (by a somewhat arbitrary amount) "because of competition" or "for the sake of rate stability" before rates are filed or used. A more realistic approach to trending might lead to better informed ratemaking decisions.


Trend Example 1
60
59
58
57
56
55
54
53
52
51
50
49
48
47
46
45
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
Data
Fitted Line


Trend Example 2
60
59
58
57
56
55
54
53
52
51
50
49
48
47
46
45
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
Data
Fitted Line

Trend Example 3