The Usefulness of the R^2 Statistic

by Ross Fonticella, ACAS

The Usefulness of the R^2 Statistic

Introduction:

Almost every Actuarial Department uses least square regression to fit frequency, severity, or pure premium data to determine loss trends. Many actuaries use the R^2 statistic to measure the goodness-of-fit of the trend. Actually, the R^2 statistic measures how significantly the slope of the fitted line differs from zero, which is not the same as a good fit.

In the Fall, 1991 Casualty Actuarial Society Forum, D. Lee Barclay wrote A Statistical Note On Trend Factors: The Meaning of R-Squared. Through simple graphical examples, Barclay showed that the coefficient of variation (R^2) is, by itself, a poor measure of goodness-of-fit. Barclay's numerical examples provide additional support for this argument. But, his paper did not analyze the formulas used in regression analysis.

By understanding the formulas and what they describe, we can further understand why the R^2 statistic is not a reliable measure of a good fit. This paper will analyze these formulas important to regression analysis: (1) the basic linear regression model, (2) the Analysis of Variance sum of squares formulas, and (3) the R^2 formula in terms of the sum of squares. With an understanding of these formulas and what they measure, actuaries can properly use the R^2 value to best determine the forecasted trend.

Formulas:

The Analysis of Variance (ANOVA) approach to regression analysis is based on partitioning the Total Sum of Squares into the Error Sum of Squares and Regression Sum of Squares.

(1) The basic linear regression model is stated as: $Y_i = B_0 + B_1 X_i$

where Y_i = the observed dependent variable

X_i = the independent variable in the ith trial

$\hat{Y}_i$ = the fitted dependent variable for the independent variable X_i

$\bar{Y}$ = mean $Y_i = \sum Y_i / n$

(2) Analysis of Variance (ANOVA) Approach to Regression Analysis

SSTO = Total Sum of Squares = $\sum (Y_i - \bar{Y})^2$
= Measure of the variation of the observed values around the mean.

SSE = Error Sum of Squares = $\sum(Y_i - \hat{Y}_i)^2$
= Measure of the variation of the observed values around the regression line.

SSR = Regression Sum of Squares = $\sum (\hat{Y}_i - \bar{Y})^2$
= Measure of the variation of the fitted regression values around the mean
= SSTO - SSE = Difference between Total and Error Sum of Squares.

(3) Coefficient of Determination: R^2 = (SSTO - SSE)/SSTO = SSR/SSTO.

What the ANOVA formulas measure when $R^2=1$ and $R^2 = 0$:

From the above formulas, we see the relevance of $R^2 = 1$. If all of the observed values (Y_i) fall on the fitted regression line: then $Y_i = \hat{Y}_i$, $SSE = \sum(Y_i - \hat{Y}_i)^2 = 0$, and $R^2 = 1$. Since there is no variation of the actual observations from the fitted values, the independent variable accounts for all of the variation in the observations Y_i.

Conversely, if the slope of the regression line is $B_1 = 0$, then $\hat{Y}_i = \bar{Y}$, $SSR = \sum (\hat{Y}_i - \bar{Y})^2 = 0$, and $R^2 = 0$ Because the SSR measures the variation in the fitted values around the mean, no variation tells us that all of the variation is explained by the mean. So the linear regression model does not tell us anything additional when the data is completely explained by the mean.

R^2 (SSR/SSTO) measures the proportion of the variation of the observations around the mean that is explained by the fitted regression model. The closer R^2 is to 1, the greater the degree of association between X and Y. Conversely, if all of the variation is explained by the mean, then $R^2 = 0$, but this should not mean that the data is not useful for forecasting purposes.

Numerical Examples:

We can use the numerical examples from Barclay's paper to examine the ANOVA formula values when $R^2 = 0$ and $R^2 = 1$. Example #1 will show that even when $R^2 = 0$, an appropriate forecast can be made by examining the data from the ANOVA formulas.

Barclay generates data from a normal distribution with a mean of 50 and variance 1 to get the observations in Example #1. The line of best fit has $B_0 = 49.38813$ and $B_1 = .0366667$.

Example #1	Y observed	Y fitted	Error (residuals)	Total	Regression
X	Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$	$Y_i - \bar{Y}$	$\hat{Y}_i - \bar{Y}$
1	48.746	49.425	-.0679	-0.844	-0.165
2	49.914	49.461	0.453	0.324	-0.128
3	49.246	49.498	-0.252	-0.344	-0.092
4	50.297	49.535	0.762	0.707	-0.055
5	48.455	49.571	-1.116	-1.135	-0.018
6	50.088	49.608	0.480	0.498	0.018
7	50.559	49.645	0.914	0.969	0.055
8	50.173	49.681	0.492	0.583	0.092
9	49.336	49.718	-0.382	-0.254	0.128
10	49.084	49.755	-0.671	-0.506	0.165
Sum	495.898	495.898	0.000	0.000	0.000
Mean	49.5898	49.590			
Sum of Squares			(SSE) 4.460	(SSTO) 4.571	(SSR) 0.111
$R^2 =$ 0.024					

The ANOVA formulas have these properties for a regression fit with a slope close to zero:

(1) $Y_i \approx \bar{Y}$, note the values in column Y fitted ($\hat{Y}$) are not far from $\bar{Y} = 49.590$.

(2) SSE ≈ SSTO

The analysis of variance sum of squares are:

$\text{SSTO} = \sum (Y_i - \bar{Y})^2 = 4.571$

$\text{SSE} = \sum (Y_i - \hat{Y}_i)^2 = 4.460$

$\text{SSR} = \sum (\hat{Y}_i - \bar{Y})^2 = 0.111$

The variation around the regression line (SSE) is not much better (smaller) than the total variation (SSTO).

(3) $R^2 = (\text{SSTO} - \text{SSE})/\ \text{SSTO} = \text{SSR}\ /\ \text{SSTO}$
$= (4.571 - 4.460)/\ 4.571 = 0.111 / 4.571 = .024$

Because the SSE is not much less than the SSTO, the R^2 value is close to 0. For SSR to be large, there needs to be a lot of variation of the fitted values around the mean. So anytime there is not a lot of variation in the data, the $R^2 \approx 0$. While this means that not much additional is explained by the fitted model, the "fit" may reasonably represent the data. And projecting with a slope of zero may be an appropriate forecast. Of course, you don't need regression to project a slope of zero, you can just forecast the mean.

In Example #2, Barclay adds 0 to the first Y observed, one to the second Y observed, two to the third, etc. The line of best fit has $B_0 = 48.38813$, and $B_1 = 1.036667$. This provides an interesting example for comparing the fit and the numerical values in the ANOVA formulas.

Example #2	Y observed	Y fitted	Error (residuals)	Total	Regression
X	Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$	$Y_i - \bar{Y}$	$\hat{Y}_i - \bar{Y}$
1	48.746	49.425	-0.679	-5.344	-4.665
2	50.914	50.461	0.453	-3.176	-3.628
3	51.246	51.498	-0.252	-2.844	-2.592
4	53.297	52.535	0.762	-0.793	-1.555
5	52.455	53.571	-1.116	-1.635	-0.518
6	55.088	54.608	0.480	0.998	0.518
7	56.559	55.645	0.914	2.469	1.555
8	57.173	56.681	0.492	3.083	2.592
9	57.336	57.718	-0.382	3.246	3.628
10	58.084	58.755	-0.671	3.994	4.665
Sum	540.898	540.898	0.000	0.000	0.000
Mean	54.0898	54.090			
Sum of Squares			(SSE) 4.460	(SSTO) 93.121	(SSR) 88.661
$R^2 =$ 0.952					

The interesting part of this example is that the residuals ($Y_i - \hat{Y}_i$) are exactly the same as in Example #1. So the SSE is the same. Recall that Linear Regression minimizes the sum of the squared residuals. Should the lines in Example #1 and Example #2 have the same fit?

Let's look at the ANOVA formulas to see the properties of a "good fit" as measured by $R^2 = 1$:

(1) $Y_i \approx \hat{Y}_i$; the fitted values ($\hat{Y}_i$ column) are close to the observed (Y_i column); a "good fit." Here we decide that $Y_i \approx \hat{Y}_i$, in favor of $Y_i \approx \bar{Y}$, because there is more variation in the observations from the mean. We choose $Y_i \approx \hat{Y}_i$, even though we have the same values for the residuals as in Example #1.

(2) SSE ≈ 0.
The analysis of variance sum of squares are:
$SSTO = \sum (Y_i - \bar{Y})^2 = 93.121$
$SSE = \sum (Y_i - \hat{Y}_i)^2 = 4.460$
$SSR = \sum (\hat{Y}_i - \bar{Y})^2 = 88.661$
The variation around the regression line (SSE) is much better (smaller) than the total variation (SSTO).

(3) $R^2 = (SSTO - SSE) / SSTO = SSR / SSTO$
$= (93.121 - 4.460)/93.121 = 88.661/93.121 = .952$

The SSE is much less than the SSTO. So a large proportion of the variation of the actual observations around the mean is being explained by the fitted line. With the SSE close to zero, most of the observations are on the fitted line. However, you will note that this is relative, because we have the same SSE as in Example #1. It is because a large proportion of the SSTO is explained by the fitted line, that we decide there is a good fit.

What does the R^2 statistic measure?

The R^2 statistic is a useful tool to determine whether or not $B_1 = 0$. For in regression, if $B_1 = 0$, there is no good reason to use the fitted line. As actuaries, we are often trying to forecast. If the slope is zero ($B_1 = 0$), then we can use the mean to forecast the fitted value.

In fact, the formula for B_1 can be written as a function of R^2:

$B_1 = [\Sigma(Y_i - \bar{Y})^2 / \Sigma(X_i - \bar{X})^2]^{1/2} r$, where $r = \pm\sqrt{R^2}$ with the sign the same as the slope.

So when $B_1=0$, then $R^2=0$; and when $R^2=0$, then $B_1=0$.

Both Example #1 and Example #2 have the same residuals, or SSE. From one perspective, each line has the same fit. The reason for the difference between the R^2 values was that in Example #2, the fitted slope is much different from zero and explains proportionally more of the larger variation in the SSTO.

In the first example, the low R^2 value would have us reject the fitted line. Should we reject the data, in favor of some other measure, like a medical CPI? I don't think so, because we can reasonably forecast that subsequent observations will be close to 49.5 (the mean). In Example #2, we get a good fit and would use $B_1 = 1.036667$. But, will the forecast of subsequent observations be any better than the forecast in Example #1? Unlikely.

The usefulness of the R^2 statistic is to measure the significance of the slope of the regression line. Since the R^2 is not a good measure of the goodness-of-fit, when the R^2 is not higher than some arbitrary benchmark, we should not just reject the data and look for other information to trend. If the slope is not significant ($R^2 = 0$) there could be a good "fit" as explained by the mean. We can see this by considering the values from the ANOVA formulas (SSE, SSR, and SSTO) which show how much of the variation is explained by the model relative to the mean. There are many other factors to be considered before accepting or rejecting the regression fit, such as patterns in the residuals. It is always useful to graph the fitted line against the observed values to look for these patterns.

Additional Formulas

The method of least squares finds values of B_0 and B_1 that minimize Q, where $Q = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - B_0 - B_1 X_i)^2$

Residuals $e_i = Y_i - \hat{Y}_i = Y_i - B_0 - B_1 X_i$

ANOVA formula relationship:

Note: The sum of the components and the sum of the squared deviations have the same relationship:

	$Y_i - \bar{Y}$	=	$\hat{Y}_i - \bar{Y}$	+	$Y_i - \hat{Y}_i$
	Total deviation	=	Deviation of fitted regression value around the mean	+	Deviation around the regression line
and	SSTO	=	SSR	+	SSE

Bibliography

John Neter and William Wasserman, Applied Linear Statistical Models, 1974.

Abraham, B.; and Ledolter, J., Statistical Methods for Forecasting, 1983.

D. Lee Barclay, A Statistical Note on Trend Factors: The Meaning of "R-Squared", Casualty Actuarial Society Forum, Fall 1991 Edition.