Some Misconceptions About R^2

By James A. Colton, M.S. and Keith M. Bower, M.S.

Abstract

The coefficient of multiple determination, commonly known as R^2, is widely used in regression analysis. Frequently, practitioners observe R^2 as a way to assess the usefulness of a particular regression model. In many instances, they are required to meet guidelines regarding "acceptable" values for R^2 (e.g. greater than 80%). As we discuss here, such arbitrary guidelines may not be appropriate in practice. This paper addresses certain areas of misunderstanding with this important statistic.

Introduction

R^2 is described in many textbooks dealing with regression analysis procedures. For further information see Hogg and Ledolter[1] (1992) or a more detailed discussion in Draper and Smith[2] (1981). Formally, we define R^2 as representing the proportion of variation in the response that is explained by the regression model. Mathematically, the general form of this relationship is:

$$R^2 = \frac{SSTO - SSE}{SSTO}$$

where SSTO is the total sum of squares in the response about the mean, and SSE is the sum of squares in the response about the regression line. Because the error in the

regression (SSE) is non-negative and cannot exceed the error in the response (SSTO), R^2 varies between 0 and 1 (i.e. 0% to 100%).

Two misconceptions regarding the use of R^2 are:

1. R^2 is very large, so the regression model is useful for predicting new observations.
2. R^2 is small, therefore no meaningful relationships exist in the data.

Misconception 1

R^2 is very large, so the regression model is useful for predicting new observations.

R^2 represents the proportion of variation in the *sample* data that is explained by the regression model. It is only an *estimate* of the proportion of variation in the *population* that is explained by the regression model. The accuracy of this estimate is greatly influenced by the technique used to select terms for the model. If the process used has a tendency to allow insignificant terms in the model (Type I error), then R^2 will have a bias toward high values. If one or more Type I errors are made in the model selection process, the resulting regression model may fit the sample data very well, producing a high R^2, yet it may not adequately fit future observations sampled from the population. Since future observations arise from the population, not the sample, a model with a high R^2 value may not necessarily be useful for prediction purposes. Note that if the model selection process

has a tendency to exclude significant terms (Type II error), then R^2 will have a bias toward low values.

Practitioners would be advised to avoid selecting a model based solely on the criterion of observing a high R^2 value. This is especially true when many terms are included in a model to fit a relatively small number of observations. For the case when no repeat runs are used, R^2 can reach the value of 1 (i.e. 100%) when we include the same number of terms as the number of dependent observations. In this situation, the practitioner would actually be modeling the error, in addition to any deterministic relationships that may exist. Such a model would therefore have little, if any, predictive ability (since error, by definition, cannot be predicted).

Including too many terms in the regression equation is called "over-fitting" the model. To prevent a misinterpretation of R^2, the results from the ANOVA table, as well as the R^2-predicted statistic should be inspected. R^2-predicted is useful for understanding the true predictive ability of a regression model.

Example 1

Consider the scatterplot of Y vs. X, in Figure 1. The Y-values were randomly generated, so there is no explicit relationship between X and Y. This is clearly illustrated by the noisy relationship exhibited between the two variables.

To illustrate how one can draw erroneous conclusions by over-fitting, consider fitting the following model to the data:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon,$$

where ε is the random error component. As is shown in the regression plot in Figure 2, R^2 is equal to 71.3%. This is interpreted as the model explaining over 71% of the variation in Y. However, this value is primarily due to the high number of variables included in the regression model, in relation to the number of observations.

The F-test in the ANOVA table (Figure 3) tests the null hypothesis, $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ vs. the alternative hypothesis, H_1: not all $\beta_j = 0$, where j = 1, 2, 3. As the p-value for this test is 0.398, there is not enough evidence to reject the null hypothesis at the $\alpha = 0.05$ significance level. In essence, we are unable to state that at least one of the regression coefficients is significantly different from zero (at the $\alpha = 0.05$ level).

A useful way to assess the ability of a model to fit future values is through R^2-predicted. Like R^2, the value of R^2-predicted varies between 0 and 1. R^2-predicted is different from R^2 however, insofar as it assesses the variability in predicting *new* observations using the regression model. For a more detailed discussion of R^2-predicted see Myers et al[3] (2002). Note from Figure 3 that the R^2-predicted value is roughly 0%, exhibiting the inability for this particular model to be used as a basis for future inferences.

The high R^2 value, along with what is visually a good fit of the regression line to the data can encourage such a misinterpretation. A model with a high R^2 value may not be useful if there are no significant effects present (as is the case here) or if the R^2-predicted value is small - especially if the model is to be used for prediction purposes.

In conclusion, since the regression coefficients are determined from the sample data, a model that is solely capable of predicting the same data that was used to create it may be of little practical use. Investigation of the parameters via the ANOVA table, and the R^2-predicted value may assist investigators in building an appropriate model.

Misconception 2

R^2 is small, therefore no meaningful relationships exist in the data.

The R^2 statistic can be small, yet one or more of the regression coefficient p-values can be statistically significant. Such a relationship between predictors and the response may be very important, even though it may not explain a large amount of variation in the response.

Example 2

Taste test experiments are sometimes used to determine final ingredient combinations for a new product. Consider the following hypothetical situation.

A taste test is conducted on two variations of a breakfast cereal. Consumers in a grocery store are randomly selected to participate in the test. Each of the 100 participants is asked to taste only one type of cereal (A or B), then score the cereal on many different attributes from 1 (extremely unfavorable) to 9 (extremely favorable).

Assume that 80% of the tasters do not adequately complete the lengthy scoring process. At some point during their evaluations, they begin providing random answers to the questions. We further assume that of this 80%, half of them tasted A, the other half B.

The scores from the participants who "correctly" completed the entire scoring process reveal that the 10% of the participants who tasted cereal A gave a lower average score than the 10% who tasted cereal B.

The random samples in this example were simulated as follows:

- Scores for the 80% of participants who answered randomly are from a normal ($\mu = 5$, $\sigma = 1$) distribution
- Scores for the 10% of participants who "correctly" evaluated cereal A are from a normal ($\mu = 4$, $\sigma = 1$) distribution
- Scores for the 10% of participants who "correctly" evaluated cereal B are from a normal ($\mu = 6$, $\sigma = 1$) distribution

A statistically significant difference may exist between the two cereals at the $\alpha = 0.05$ significance level, in spite of the low R^2 value. This is true for the simulated data, as shown in Figures 4 and 5 where the p-value (0.045) is less than α (0.05), and $R^2 = 4.1\%$.

The low R^2 value reflects the high level of variation in the scores of the tasters who randomly guessed. Regardless of the higher variation, the results from this analysis indicate (correctly) that cereal B is preferred and would likely result in increased sales over cereal A.

Many other real-world occurrences can lead to lower values of R^2, though significant effects that have practical importance may be present. Examples include situations involving high levels of measurement error, or wide variation reflecting consumer behavior (as in economic data).

In addition, as with misconception 1, it is important to note that R^2 is reflecting variation solely obtained from the sampled data. If the data exhibit an inflated error that does not represent the true level of variation present in the overall population (e.g. due to incorrect sampling techniques), R^2 can be low, while meaningful relationships may still exist.

Conclusion

We have dealt with two instances in which the assessment of R^2 may lead to invalid conclusions. It is the responsibility of the analyst to recognize when to disregard a high R^2 value due to "over-fitting" the model, and to disregard a low R^2 value due to large error in the sampled data. Furthermore, these two misconceptions point out why

establishing a threshold or cut-off point for an "acceptable" value of R^2 across all applications is inappropriate.

Figure 1

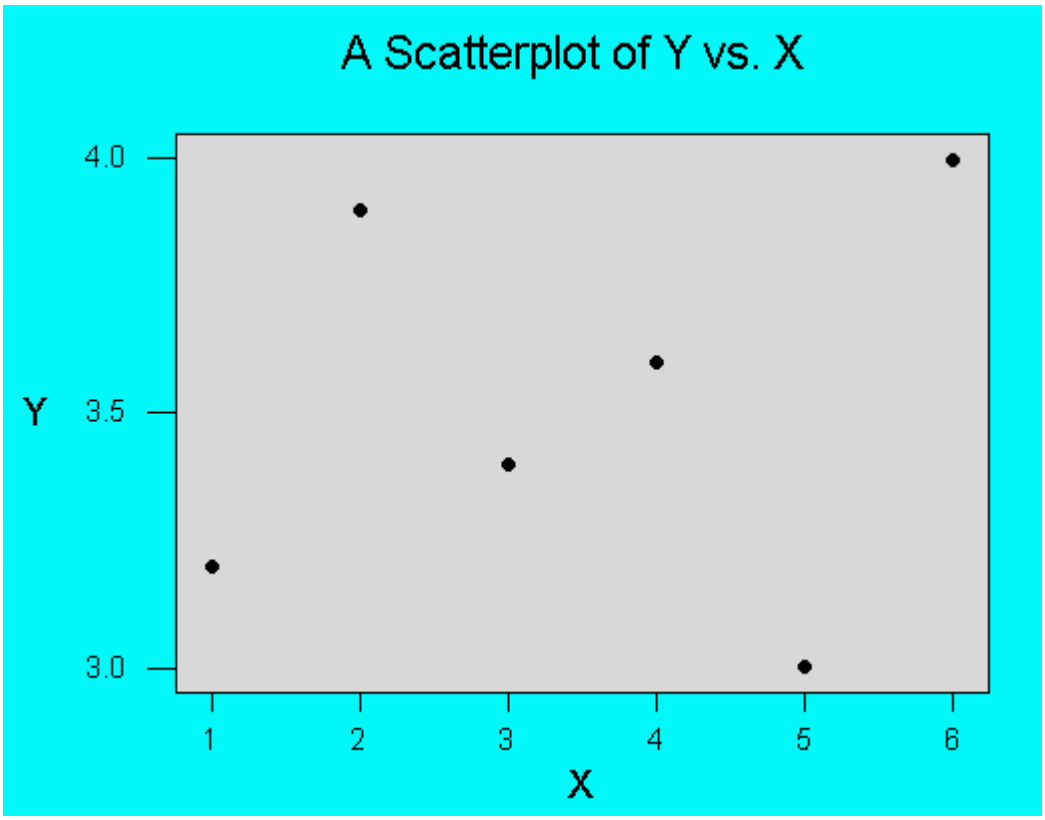

Figure 2

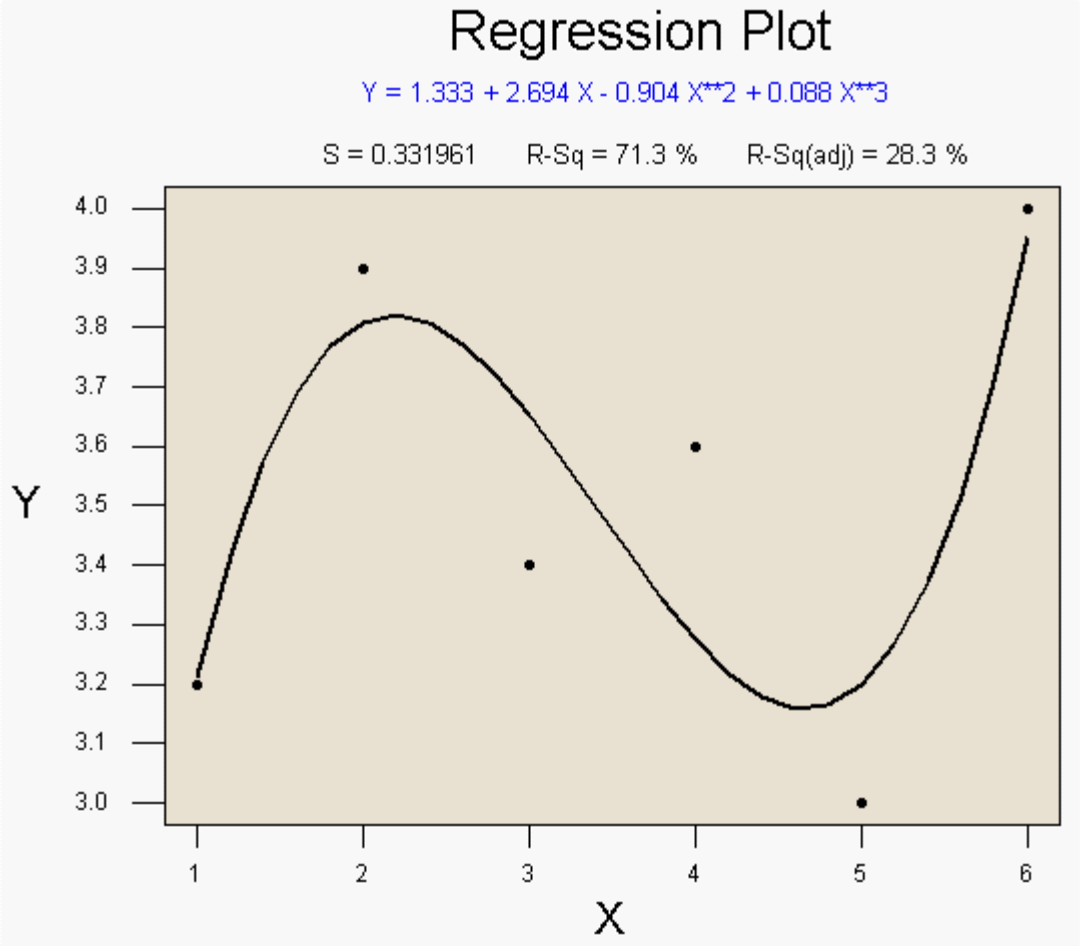

Figure 3

Regression Analysis: Y versus X, X^2, X^3

The regression equation is
Y = 1.33 + 2.69 X - 0.904 X^2 + 0.0880 X^3

Predictor	Coef	SE Coef	T	P
Constant	1.333	1.197	1.11	0.381
X	2.694	1.364	1.98	0.187
X^2	-0.9040	0.4364	-2.07	0.174
X^3	0.08796	0.04124	2.13	0.167

S = 0.3320 R-Sq = 71.3% R-Sq(adj) = 28.3%
PRESS = 2.32537 R-Sq(pred) = 0.00%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	0.5479	0.1826	1.66	0.398
Residual Error	2	0.2204	0.1102		
Total	5	0.7683			

Figure 4

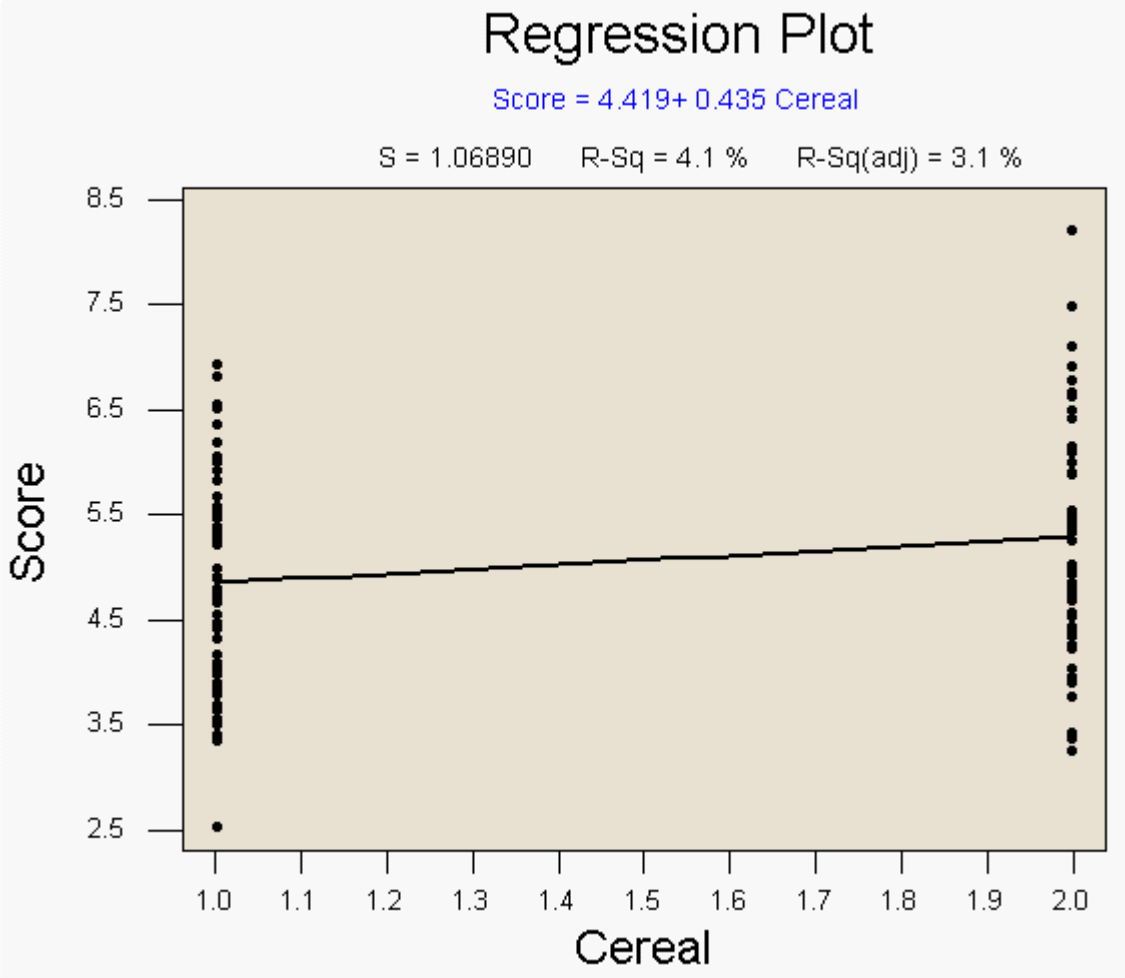

Figure 5

Regression Analysis: Score versus Cereal

The regression equation is
Score = 4.42 + 0.435 Cereal

Predictor	Coef	SE Coef	T	P
Constant	4.4187	0.3380	13.07	0.000
Cereal	0.4349	0.2138	2.03	0.045

S = 1.069 R-Sq = 4.1% R-Sq(adj) = 3.1%
PRESS = 116.587 R-Sq(pred) = 0.10%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	4.728	4.728	4.14	0.045
Residual Error	98	111.970	1.143		
Total	99	116.698			

References

1. Hogg, R.V., Ledolter J. 1992. *Applied Statistics for Engineers and Physical Scientists*: Macmillan
2. Draper, N.R., Smith, H. 1981. *Applied Regression Analysis*: Wiley
3. Myers, R.H., Montgomery, D.C., Vining, G.G. 2002. *Generalized Linear Models: with applications in engineering and the sciences*: Wiley

James A. Colton has an M.S. in Quality and Applied Statistics from the Rochester Institute of Technology and an M.S. in Statistics from The Ohio State University. Keith M. Bower has an M.S. in Quality Management and Productivity from The University of Iowa.

Both are Technical Training Specialists with Minitab Inc.