Contributions to Methodology and Statistics
A. Ferligoj and A. Kramberger (Editors)
Metodološki zvezki, 10, Ljubljana: FDV 1995

Some Properties of R^2 in Ordinary Least Squares Regression

Janez Stare[1]

Abstract

A lot has been written about R^2 in order to discourage or at least warn against its use. Many of the papers address the use and definition of R^2 in generalized linear models where its different interpretations lead to different statistics, but quite a few has also been said about the drawbacks of R^2 in ordinary least square regression. The authors usually question its use as a measure of goodness of fit, complain that R^2 can not be 1 when there are replicated responses or show that its value depends on the range of independent variables which makes it difficult to estimate its population value. Often modifications of R^2 or alternative statistics are proposed. All these arguments raise doubt about the adequacy of using R^2. We argue that this doubt is usually not justified and that there is no great danger in using R^2 in ordinary least squares regression.

1 Introduction

The coefficient of determination R^2 is widely used as a measure of predictive power of linear regression models. The main reason for this is the interpretation of R^2 as a proportion of variation of the dependent variable explained or accounted for by the model. In other words, R^2 tells us how well our model explains the occurrence of different values of the outcome.

Although R^2 has some nice properties in OLSR, much has been written to dispraise its use even in this setting. It is obvious that some problems arise from a different understanding of the concept of *goodness-of-fit*. The distinction between the concepts of *explained variation* and the *goodness-of-fit* is also not always clear, which adds to the differences in views about the usefulness of R^2. The paper by Korn and Simon (1991) makes a great contribution to the clarification of this misunderstanding. They demonstrate that explained variation measures both the

[1] Institute for Biomedical Informatics, Medical Faculty, P.O.Box 18, 61105 Ljubljana, Slovenia

explained risk and the *goodness-of-fit* of a model. By goodness-of-fit they understand the consistency of the model with the data, and the explained risk is a measure of how much better predictions are when one uses covariates compared to not using them.

2 Definition and interpretation of R^2 in ordinary least squares regression

We review here some well-known results without giving any proofs.

Let $x_1, x_2, \ldots, x_p$ and y denote $p+$ variables. For any given set of values, $x_{10}, x_{20,} \ldots, x_{p0}$ say, we may be interested in the conditional expectation (mean) of y, denoted by $E(y|x_{10},, x_{p0})$. When such conditional expectation is defined for all x-values, the function $E(y|x_1, \ldots, x_p)$ is called the *regression curve* of y on $x_1, x_2, \ldots, x_p$. When this function takes the form

$$E(y|x_1, \ldots, x_p) = \alpha + \sum_{i=1}^{p} \beta_i x_i \,,$$

we talk about *linear regression*. The conditional variance $\mathrm{var}(y|x_1, \ldots, x_p)$ shall be denoted by $\sigma^2{}_{y.1\ldots p}$.

The population $\mathbf{R}^2$ is defined by[2]

$$1 - \mathbf{R}^2 = \frac{\sigma^2_{y.1\ldots p}}{\sigma^2_y}.$$

The positive square root of $\mathbf{R}^2$ is called the *multiple correlation coefficient* and is denoted by $\mathbf{R}$. When we want to emphasize that we are talking about the correlation between y and $x_1, x_2, \ldots, x_p$, we write $\mathrm{R}_{y(1\ldots p)}$. It can be shown that $\mathbf{R}$ is indeed a correlation coefficient (see for example Stuart and Ord, 1991).

Most often we are not interested in the population $\mathbf{R}^2$ but in its sample analogue, denoted by R^2. It can be defined in the same way, replacing population variances with sample estimates. However, another approach to its definition will

[2] We follow the notation used by Stuart and Ord (1991). Thus, bold faced $\mathbf{R}^2$ denotes the population value while R^2 stands for the sample estimate.

be given, an approach that is much more common and more natural (see for example Draper and Smith, 1981).

Let $y_i, i = 1,\ldots,n$, denote the observed values of the dependent variable, its mean $\bar{y}$ and $\hat{y}_i$ the predicted values. The situation is depicted in Figure 1.

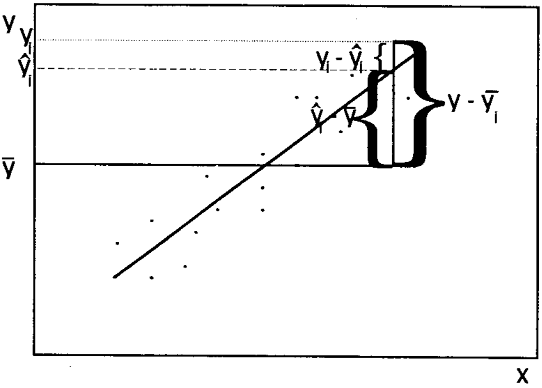


Figure 1.

We can write

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}).$$

Squaring both sides and summing over i gives

$$\sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2, \qquad (2.1)$$

where the cross-product term is omitted since it is equal to 0. The left-hand side of equation (2.1) is the sum of squares of deviations of the observed values from the mean, usually called 'SS about the mean'. On the right-hand side of this equation we have the sum of squares of deviations of observed values from predicted values $(y_i - \hat{y}_i)$ and the sum of squares of deviations of predicted values from the mean $(\hat{y}_i - \bar{y})$. Quantities $(y_i - \hat{y}_i)$ are usually called ***residuals***. Equation (2.1) can then be expressed as

SS about the mean = SS about regression + SS due to regression.

We see that SS about the mean is divided in two parts. One would expect from a good regression model that residuals were small, that is, that SS due to regression was much greater than SS about regression. In other words, one would like to have the ratio (SS due to regression / SS about the mean) as close to 1 as possible. Since SS about the mean can be regarded as the total variability observed in the dependent variable and SS due to regression as the amount of this variability that is explained, the ratio

$$R^2 = \frac{\text{SS due to regression}}{\text{SS about the mean}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \tag{2.2}$$

represents the proportion of the total variability in the dependent variable that is *explained* by the independent variables. R^2 is often called *coefficient of determination*. It can be shown that R^2 is the square of the sample correlation coefficient between y and the best fitting linear combination of $x_1, x_2, \ldots, x_p$.

Another definition of R^2 is often given

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}. \tag{2.3}$$

In ordinary least squares regression (OLSR), the two definitions are equivalent because of property (2.1). Kvalseth (1985) lists other definitions and discusses their properties in nonlinear regression. He also gives a list of general properties that R^2 should possess. Based on this list, he decides on definition (2.3) as being the most useful for more general models.

What is so appealing about R^2? It seems that the following properties make it so useful:

- it has an intuitively clear interpretation,
- it is a number that can be easily calculated when the model is fitted,
- it is invariant to units of measurement,
- it lies between 0 and 1,
- it becomes larger when the model 'fits better'.

It should be noted that the value of R^2 does not depend only on the distances between predicted and observed values but also on the variation of the outcome variable. So anything that influences this variation also influences the value of R^2. This is evident from the definition (2.3).

The interpretation and use of R^2 has been extensively discussed in the literature (some examples are in Crocker, 1972; Barrett, 1974; Draper and Smith, 1981; Ranney and Thigpen, 1981; Healy, 1984; Kvalseth, 1985; Helland, 1987;

Willett and Singer, 1988; Nagelkerke, 1991; Scott and Wild, 1991). We return to some of these later. While the distribution of R^2 and the inference about its population value are not the issues of main interest here, two points must be stated:

- with increasing n, R^2 tends to

$$\frac{\beta \mathbf{S}_x \beta}{\beta \mathbf{S}_x \beta + \sigma^2_{y.1\ldots p}}, \tag{2.4}$$

where $\mathbf{S}_x$ is the sample covariance matrix for the independent variables (Helland, 1987). Thus, the value of R^2 depends on the variation among independent variables.

- under the null hypothesis $\mathrm{R}^2 = 0$, the expected value of R^2 can be shown to be

$$E(R^2) = \frac{p}{n-1},$$

where p is the number of independent variables and n is the sample size. This means that we can expect values of sample R^2 greater than 0 even if its population value is 0. This property of R^2 is another reason for requiring large n/p ratios in regression analysis.

Because of this second property, a modified R^2, called *adjusted-R^2* is sometimes used. However, apart from being smaller than usual statistics, there is no other reason for its use.

The following points were raised against R^2:

- it can not be 1 when there are replicated responses,
- it diminishes with replicated responses,
- it depends on the slope of the regression plane,
- it depends on the range of independent variables.

3 Analysis of some properties of R^2 in ordinary least squares regression

It was stated in the introduction that criticism has often been addressed to the use of R^2 even in ordinary regression. We will show here that this criticism is mostly unjustified. The points raised against R^2 will be discussed using examples from bivariate linear regression which can readily be generalized to the multivariate case.

3.1 The slope

Figures 3.1.a and 3.1.b illustrate the dependence of R^2 on the slope of the line. The two lines have different slopes, but the points are at the same distance from the line. We see that there is a big difference in R^2, which is clearly due to greater variation of the dependent variable in the second case. Some authors (Barrett, 1974) argue that a measure of goodness-of-fit should be the same in both cases, since the precision of prediction is the same. This is true if we measure this precision just by the distances of the points from the curve. In this sense even a model with $R^2 = 0$ could have the same precision. If we take into account the relative gain from going from the null model to the fitted model then this gain is much greater in the second case, and this is reflected in R^2. The second model is clearly more capable of distinguishing the differing outcomes and this is an important feature in judging the quality of the model. Alternatively, following Korn and Simon (1993), we can say that changing the slope of the line changes the explained risk while goodness-of-fit is not affected.

Figure 3.1.a.

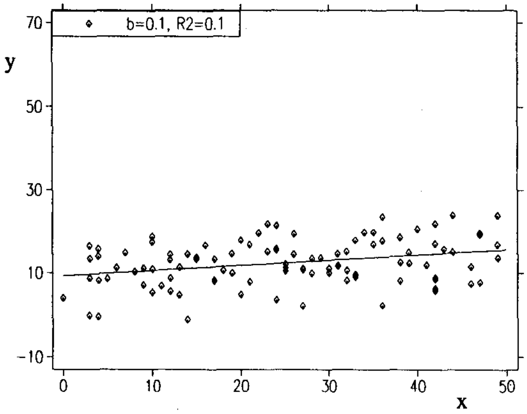
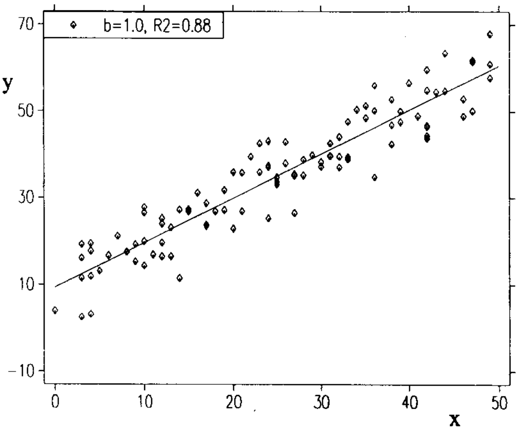


Figure 3.1.b.



3.2 R^2 cannot be 1 when there are replicated responses

It is of course clear that if we have different outcomes for the same values of the independent variable there will always remain some unexplained variation. This has been a motive for different attempts to modify R^2 for such cases in order to have a statistic that has 1 as its upper limit. All these attempts have their drawbacks which we will not discuss here (see for example Chang and Afifi, 1987). Figures 3.2.a and 3.2.b show two models for which these modified measures would be 1 but R^2 is different. The reason is that R^2 takes into account the spread of the outcomes around the line, while the modified measures don't.

The possibility of R^2 reaching 1 actually does not depend on the replicated responses. If the true R^2 is 1 then replicated responses are all equal, while if the true R^2 is not 1 then not having replicated responses does not change anything. If we are sampling from the data that follow a linear model with normally distributed error, the probability of obtaining points that lie on a straight line is exactly the same as getting equal responses with repeated measurements. In both cases we must get exactly the means of the underlying normal distribution. Modified statistics can only be looked at as statistics that give *additional* information, not as a *replacement* for R^2. It is confusing to use one statistic when there are replicated responses and another when there are no replicated responses.

Behaviour of R^2 with replicated responses

Figure 3.2.a

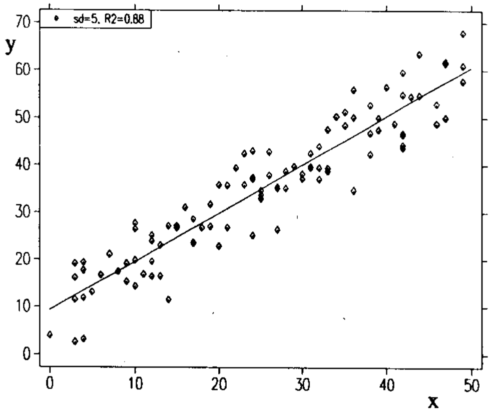
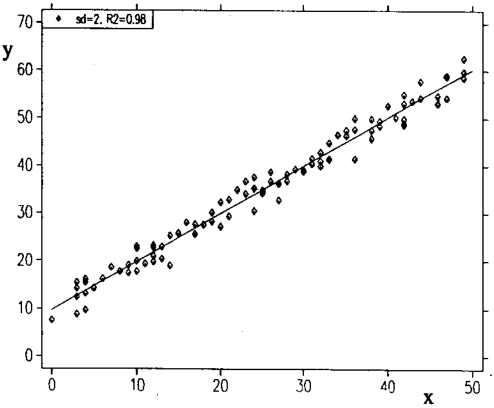


Figure 3.2.b.



3.3 The diminishing of R^2 with replicated responses

This is actually not true and we could stop talking about this point here. But as it seems that there is some confusion about this, it is probably good to say some words on this topic. We don't know if the confusion existed before Draper's paper in 1984, but it was certainly amplified after that. The paper was for example cited in the same journal in the same year by Healy (1984), who took the assertions from Draper's paper for granted. The fact that Draper actually corrected his mistake in 1985 seems to have been overlooked even after many years (see for example Scott and Wild, 1991).

The non-dependence of R^2 on the replicated responses can be shown in different ways:

- first, it can be seen from the formula (2.4) for the asymptotic value of R^2 that with a given range of independent variables, the value of R^2 depends only on the variance around the fitted line .

- second, the example illustrated in Table 3.1 provides empirical proof that Draper's argument was wrong. The letter k in the table denotes the number of different values of x and N stands for the number of cases. The independent variable x was randomly assigned integer values between 0 and 50. The dependent variable y was then generated in the following way

$$y = x + 10 + e,$$

where the error term e was taken to follow normal distribution $N(0,5)$. It can be seen that even with N-k dramatically increasing, R^2 does not change apart from random variation.

Table 3.1. The dependence of R^2 on replicated responses.

	N=100 k=43	N=200 k=50	N=1200 k=50	N=5000 k=50	N=10000 k=50
R^2	.8851	.8796	.8860	.8933	.8913

- third, simple algebraic consideration shows that replicated responses have no effect on R^2. Suppose we estimated R^2 using two different samples in such a way that x takes the same values in both samples. For the sake of argument assume that the estimates are equal. Let us denote the corresponding SS due to regression by SR_i and SS about the mean by SM_i $(i = 1,2)$. Then we have

$$R^2 = \frac{\mathrm{SR}_1}{\mathrm{SM}_1} = \frac{\mathrm{SR}_2}{\mathrm{SM}_2}.$$

It follows that

$$\mathrm{SR}_1 \cdot \mathrm{SM}_2 = \mathrm{SR}_2 \cdot \mathrm{SM}_1. \tag{3.1}$$

If we take both samples together we get the following estimate for R^2

$$R^2 = \frac{\mathrm{SR}_1 + \mathrm{SR}_2}{\mathrm{SM}_1 + \mathrm{SM}_2}.$$

This is clearly an estimate that we get after adding replicated responses to the first sample. When is this estimate equal to the estimates from the separate samples? For this to be true we should have

$$\frac{\mathrm{SR}_1 + \mathrm{SR}_2}{\mathrm{SM}_1 + \mathrm{SM}_2} = \frac{\mathrm{SR}_1}{\mathrm{SM}_1},$$

and from this

$$\mathrm{SR}_1 \cdot \mathrm{SM}_1 + \mathrm{SR}_1 \cdot \mathrm{SM}_2 = \mathrm{SR}_1 \cdot \mathrm{SM}_1 + \mathrm{SR}_2 \cdot \mathrm{SM}_1.$$

After cancelling $\mathrm{SR}_1 \cdot \mathrm{SM}_1$ we get

$$\mathrm{SR}_1 \cdot \mathrm{SM}_2 = \mathrm{SR}_2 \cdot \mathrm{SM}_1,$$

which is always true as we know from (3.1).

3.4 The dependence of R^2 on the range of the independent variables

This is exemplified in Figure 3.3. The first figure represents a random sample of a population for which the true model is

$$y = 5 + 1.25x + e,$$

with e distributed as N(0,9). In the next figure, values of x between 4 and 17 are not used, and in the third figure the values outside the interval (4,17) are not used. The effect on R^2 is large. The reason for this is obvious: such sampling effects the variance of the outcome. Now, it is easier to accept the increase in R^2 from Figure 3.3.c to Figure 3.3.a; in Figure 3.3.a we have more evidence to support the calculated curve, our prediction is relatively better since the points on the edges of the curve are furthest from the mean and contribute more to the variance of the outcome than the points in the middle.

But it is hard to accept that less evidence in the second figure gives a bigger R^2. Beside the fact that only the further points were used, one should also stress that drawing a straight line through the whole range of the independent variable is actually an extrapolation of the regression curve, which is bad statistical practice. Of course, if there is good reason to believe that the model actually holds in the whole range, this property of R^2 is annoying and one should be careful in choosing the values of independent variables if one can do so.

If we want to estimate the real population value of R^2 there is no alternative to taking random samples (Helland, 1987). The dependence of R^2 on the distribution of covariates is also a property which makes this statistics often inappropriate for comparing models built on different data sets. Unless we are sure that the two samples represent the same population, comparison of models based on R^2 is not suitable.

Dependence of R^2 on the range

Figure 3.3.a.

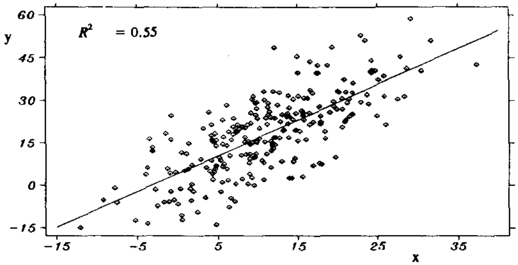


Figure 3.3.b.

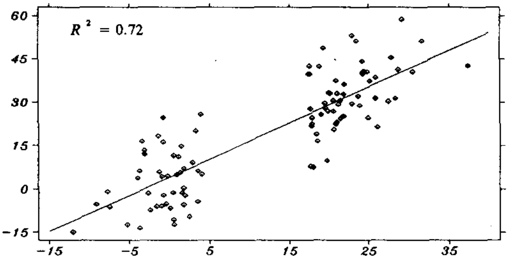
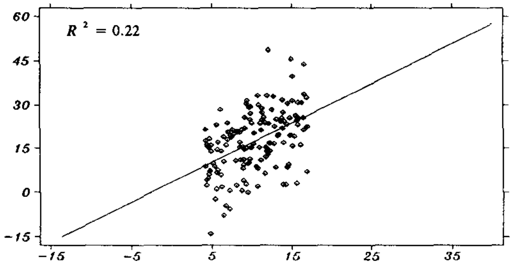


Figure 3.3.c.



References

[1] Barrett J.P. (1974): The coefficient of determination - some limitations. *The American Statistician*; **28**, 19–20.

[2] Chang P.C. and Afifi A.A. (1987): Goodness-of-fit statistics for general linear regression equations in the presence of replicated responses. *The American Statistician*, **41**, 195-99.

[3] Crocker D.C. (1972): Some interpretations of the multiple correlation coefficient. *The American Statistician*, **26**, 31–3.

[4] Draper N.R. (1984): The Box-Wetz criterion versus R2. *J R Statist Soc A*, 147, **Part 1**: 101-3.

[5] Draper N.R. (1985): Corrections. The Box-Wetz criterion versus R2. *J R Statist Soc A*, 148, **Part 4,** 357.

[6] Draper N.R. and Smith H. (1981): *Applied regression analysis, 2nd ed.* New York: Wiley.

[7] Healy M.J.R. (1984): The use of R2 as a measure of goodness of fit. *J R Statist Soc A* , 147, **Part 4**: 608–9.

[8] Helland I.S. (1987): On the interpretation and use of R2 in regression analysis. *Biometrics*, **43**, 61–9.

[9] Korn E.L. and Simon R. (1993): Explained residual variation, explained risk, and goodness of fit. *The American Statistician*, **45**, 201–6.

[10] Kvalseth T.O. (1985): Cautionary note about R2. *The American Statistician*, **39**, 279–85.

[11] Nagelkerke N.J.D. (1991): A note on a general definition of the coefficient of determination. *Biometrika*, **78**, 691–2.

[12] Ranney G.B. and Thigpen C.C. (1981): The sample coefficient of determination in simple linear regression. *The American Statistician*, **35**, 152 –3.

[13] Scott A. and Wild C. (1991): Transformations and R2. *The American Statistician*, **45**, 127–9.

[14] Stuart A. and Ord J.K. (1991): *Kendall's advanced theory of statistics*, 5th ed., Vol. 2: Classical inference and relationship. London: Edward Arnold.

[15] Willett J.B. and Singer J.D. (1988): Another cautionary note about R2: its use in weighted least-squares regression analysis. *The American Statistician*, **42**, 236–8.