The acceptable R-square in empirical modelling for social science research

Ozili, Peterson K

2023

The Acceptable R-square in Empirical Modelling for Social Science Research

Peterson K. Ozili

Abstract

This article examines the acceptable R-square in social science empirical modelling with particular focus on why a low R-square model is acceptable in empirical social science research. The paper shows that a low R-square model is not necessarily bad. This is because the goal of most social science research modelling is not to predict human behaviour. Rather, the goal is often to assess whether specific predictors or explanatory variables have a significant effect on the dependent variable. Therefore, a low R-square of at least 0.1 (or 10 percent) is acceptable on the condition that some or most of the predictors or explanatory variables are statistically significant. If this condition is not met, the low R-square model cannot be accepted. A high R-square model is also acceptable provided that there is no spurious causation in the model and there is no multi-collinearity among the explanatory variables.

Keywords: R-square, low R-square, social science, research, empirical model, modelling, regression.

January 2023

Published in Book: ***Social Research Methodology and Publishing Results***

1. Introduction

This paper examines the acceptable R-square in empirical social science research. It focuses on why a low R-square model is acceptable in empirical social science research.

As a general principle, an econometric model is considered to have a high predictive power if the model has a high R-square or adjusted R-square (Gujarati, Porter and Gunasekar, 2012). This general principle often gives the scientist some confidence that the explanatory variables in the model are good predictors of the dependent variable (Hill, Griffiths and Lim, 2018).

Many social scientists, who follow this principle, are often excited when their models report a high R-square and they get worried when their models report a very low R-square. Their worry is further amplified when they learn that statisticians and scientists in the pure sciences will dismiss a model as "weak", "unreliable" and "lacking a predictive power" if the reported R-square of the model is below 0.6 (or 60 percent when expressed in percentage). In this paper, I address this issue and show that empirical modelling in social science has a different purpose compared to empirical modelling in the pure science.

The rest of the paper is structured as follows. Section 2 discuss the imperfect nature of social science. Section 3 highlights the different range of R-square. Section 4 presents the conclusion.

2. Literature review

There is adequate literature about the R-squared. The literature about R-squared shows some of its applications. Miles (2005) showed that the R-squared and the adjusted R-squared statistics are derived from analyses based on the general linear model (e.g., regression, ANOVA), and they represent the proportion of variance in the outcome variable which is explained by the predictor variables in the sample (R-squared) and an estimate in the population (adjusted R-squared). Hagle and Mitchell (1992) suggested a refinement to the R-squared called the pseudo R-squared. They suggest that the corrected Aldrich-Nelson pseudo R-squared is a good estimate of the R-squared of a regression model because of its smaller standard deviations and range of errors, and its smaller error of regression. They also point out that the Aldrich-Nelson correction to the R-squared is more robust when the assumption of normality is violated. However, they cautioned that the pseudo R-squared should be used with caution because even a good summary measure can be misinterpreted; therefore, it was suggested that the pseudo R-squared should be used in conjunction with other measures of model performance. Chicco et al (2021) suggested that the use of the R-squared statistic as a standard metric to evaluate regression analyses is popular in any scientific domain. This is because the coefficient of determination (or R-squared) is more informative and truthful than other goodness of fit measures. Cameron and Windmeijer (1997) showed that R-squared type goodness-of-fit summary statistics have been constructed for linear models using a variety of methods. They propose an R-squared measure of goodness of fit for the class of exponential family

regression models, which includes logit, probit, Poisson, geometric, gamma, and exponential. They defined the R-squared as the proportionate reduction in uncertainty, measured by Kullback-Leibler divergence, due to the inclusion of regressors. They also show that, under further conditions concerning the conditional mean function, the R-squared can also be interpreted as the fraction of uncertainty explained by the fitted model. Gelman et al (2019) argued that the usual definition of the R-squared statistic (variance of the predicted values divided by the variance of the data) has a problem for Bayesian fits, as the numerator can be larger than the denominator. They proposed an alternative definition similar to one that has appeared in the survival analysis literature. The definition they propose defined the R-squared as the variance of the predicted values divided by the variance of predicted values plus the expected variance of the errors.

Hagquist and Stenbeck (1998) examined the importance of the R-square as a measure of goodness of fit in regression analysis. They argued that the utility of goodness of fit measures depends on whether the analysis focuses on explaining the outcome (model orientation) or explaining the effect(s) of some regressor(s) on the outcome (factor orientation). They further argued that in some situations a decisive goodness of fit test statistic exists and is a central tool in the analysis. In other situations, where the goodness of fit measure is not a test statistic but a descriptive measure, it can be used as a heuristic device along with other evidence whenever appropriate. They also argued that the availability of goodness of fit test statistics depends on whether the variability in the observations is restricted, as in table analysis, or whether it is unrestricted, as in OLS and logistic regression on individual data.

Some scholars consider the R-squared to be of limited importance (e.g., King, 1986; King, 1990). Others consider the R-squared to be very useful (Lewis-Beck, and Lweis-Beck, 2015). Lewis-Beck and Skalaban (1990) sounded a word of caution about using R-squared. They argued that a researcher should be careful to recognize the limitations of the R-squared as a measure of goodness of fit, and that one must be extremely cautious in interpreting the R-squared value for an estimation and particularly in comparing R-squared values for models that have been estimated with different data sets. They also argued that the R-squared measures nothing of serious importance because it measures nothing of importance in most practical political science situations and the interpretation of the R-square adds little meaning to political analyses.

Onyutha (2020) criticized the R-squared and argued that the most extensively applied goodness-of-fit measure for assessing performance of regression models is the R-squared or coefficient of determination. The author showed that although a high R-squared tends to be associated with an efficient model, the R-squared has been cited to have no importance in the classical model of regression because it does not give any information on the model residuals. The author also argued that a very poor model fit can yield a high R-squared, and more importantly, regressing X on Y yields an R-squared which is the same as that if Y is regressed on X, thereby invalidating its use as a coefficient of determination. Figueiredo Filho et al (2011) argued that no substantive meaning can be drawn from the R-squared statistic. They argued that the R-squared cannot help us to make causal claims about the relationship between the independent variables and the dependent variable,

and that the R-squared does not assist us regarding omitted variable bias. They further argued that the R-squared does not inform us if X1 is strongly correlated with X2 (collinearity problems in the data).

Maydeu-Olivares and Garcia-Forero (2010) argued that the goodness of fit of many models cannot be assessed using summary statistics such as proportions or covariance. They argued that, in the context of linear regression and related models, the R-square is sometimes described as a goodness of fit statistic, but if a linear regression model can fit the data perfectly, the R-squared will be zero if the slope is zero. Ijomah (2019) showed that the R-squared is the single most extensively used measure of goodness of fit for regression models, and it measures the proportion of variation in the dependent variable explained by the predictors included in the model. But it is widely misused as the square of correlation coefficient, and this has led to poor interpretation of research reports in regression model. CSCU (2005) argued that one pitfall of R-squared is that it can only increase as predictors are added to the regression model. This increase is artificial when predictors are not actually improving the model's fit. They show that to remedy this problem, the adjusted R-squared incorporates the model's degrees of freedom. They show that the adjusted R- squared will decrease as predictors are added if the increase in model fit does not make up for the loss of degrees of freedom; likewise, it will increase as predictors are added if the increase in model fit is worthwhile.

Reisinger (1997) found that the R-squared gets smaller as the sample size increases and the number of regressors decreases in a study. The author also found that time-series studies achieve higher values for R-squared than cross-sectional studies. The author further found that studies with secondary data achieve higher values for R-squared than studies with primary data. Cornell and Berger (1987) argued that the R-squared is one of the most widely used statistics and it is determined by calculating the proportion of the total variation in goodness-of-fit of the equation. However, they pointed out that the R-squared value is affected by several factors, some of which are associated more closely with the data collection scheme or the experimental design than with how close the regression equation fits the observations. They argued that the design factors are: the range of values of the independent variable (X), the arrangement of X values within the range, the number of replicate observations (Y), and the variation among the Y values at each value of X. Ferligoj and Kramberger (1995) showed that, even though scholars usually question the use of R-squared as a measure of goodness of fit in ordinary least square regression, there is no great danger in using R-squared in ordinary least squares regression. Spiess and Neumeyer (2010) argued that the R-squared is inappropriate when used for assessing model performance in certain nonlinear models especially in pharmacological and biochemical scientific research.

3. The imperfect nature of social science

In the pure science, models should have a high predictive power or a high R-squared. This is because researchers in the pure sciences deal with molecules, materials, objects or atoms whose properties are known and whose behaviour are predictable and do not change over time. As a result, it is reasonable to expect a high R-squared in the models used in the pure science. In contrast, the social sciences deal with human behaviour or human relationship that is subject to change from time to time. Human behaviour may change due to individual self-interest, group dynamics, feelings and other factors. For this reason, it is difficult to accurately predict human behaviour in the social sciences; therefore, the modelling of human behaviour will be an imperfect science and it will be difficult for a single model to capture all the factors that predict human behaviour at a given time. And even if it is possible to include all the explanatory variables that explain human behaviour into the model, some of the included explanatory variables may have a weak or non-linear relationship with the dependent variable thereby weakening the R-squared goodness-of-fit of the model.

4. R-squared categories

4.1. Negative R-squared

In social science empirical modelling, a univariate linear model that reports a negative R-squared should be rejected. Similarly, a multivariate linear model that reports a negative adjusted R-squared should be rejected. The reason is because the model shows that the explanatory variables do not predict the changes in the dependent variable.

A negative R-squared or negative adjusted R-squared value means that the reported predictive power of the model is less accurate than the average value of the dataset over time. It also means that the model is predicting worse than the mean of the dataset. The implication is that the explanatory variables do not predict the specific human behaviour or social norm being estimated. For example, assume the profit-before-tax (PBT) of a firm depends on its cost ratio (COST), the prevailing rate of inflation (INFLATION) and the revenue ratio (REVENUE). Estimating the multivariate regression model using the data set below and using the ordinary least square regression method would yield a negative R-squared and a negative adjusted R-squared.

Table 1. Hypothetical Data

Year	PBT (%)	Cost (%)	Inflation (%)	Revenue (%)	Taxes (%)
2001	16	25	2	54	15
2002	14	22	2	38	15
2003	12	34	3	47	14.5
2004	17	27	4	49	16
2005	16.5	28	3	51	15
2006	18	30	2.5	50	20
2007	14.5	31	4.5	49	21
2008	11	34	3.8	54	10
2009	19	35.6	2.9	37	14
2010	15	29	3.4	50	15

Table 2. OLS Result for Negative R^2 and adjusted R^2

Dependent Variable: PBT
Method: Least Squares
Date: 11/17/22 Time: 12:21
Sample: 2001 2010
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
COST	0.317369	0.239697	1.324042	0.2271
INFLATION	-0.912513	1.666351	-0.547612	0.6010
REVENUE	0.179281	0.136803	1.310501	0.2314

R-squared	-0.556121	Mean dependent var	15.30000
Adjusted R-squared	-1.000726	S.D. dependent var	2.529822
S.E. of regression	3.578358	Akaike info criterion	5.631010
Sum squared resid	89.63254	Schwarz criterion	5.721786
Log likelihood	-25.15505	Hannan-Quinn criter.	5.531430
Durbin-Watson stat	2.886551		

4.2. R-squared between 0 and 0.09

A R-squared between 0 and 0.09 (or between 0% to 9%) is too low for an empirical model in social science research. This range of R-squared is not acceptable. It should be rejected. A social science researcher who is faced with a low R-squared within this range can increase the R-squared by doing one of these: (i) replace the explanatory variables with a new set of explanatory variables, (ii) introduce additional explanatory variables together with the existing explanatory variables in the model, (iii) change the entire dataset, (iv) change the model estimation method, (v) change the dependent variable, (vi) change the structural form of the model, (vii) remove the highly correlated explanatory variables, or (viii) combine the highly correlated variables.

4.3. R-squared between 0.10 and 0.50

A R-squared that is between 0.10 and 0.50 (or between 10 percent and 50 percent when expressed in percentage) is acceptable in social science research only when some or most of the explanatory variables are statistically significant. For example, assume the profit-before-tax (PBT) of a firm depends on its cost ratio (COST), the prevailing rate of inflation (INFLATION), and the tax rate (TAX). Estimating the multivariate regression model using the data set below and using the ordinary least square regression method yields an of R-squared of 0.106. A model with a R-squared that is between 0.10 and 0.50 is good provided that some or most of the explanatory variables are statistically significant. In table 3, the COST and TAX variables are statistically significant at the 10% and 5% level. However, a model with a R-squared that is between 0.10 and 0.50 must be rejected if all the explanatory variables in the model are statistically insignificant.

Table 3. OLS Result for Negative R^2 between 0.1 and 0.5

Dependent Variable: PBT
Method: Least Squares
Date: 11/17/22 Time: 12:51
Sample: 2001 2010
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
COST	0.297996	0.154547	1.928190	0.0952
INFLATION	-1.182598	1.249790	-0.946237	0.3755
TAX	0.647678	0.226518	2.859280	0.0244

R-squared	0.106101	Mean dependent var	15.30000
Adjusted R-squared	-0.149298	S.D. dependent var	2.529822
S.E. of regression	2.712104	Akaike info criterion	5.076652
Sum squared resid	51.48857	Schwarz criterion	5.167427
Log likelihood	-22.38326	Hannan-Quinn criter.	4.977071
Durbin-Watson stat	2.110080		

4.4. R-squared between 0.51 and 0.99

A R-squared between 0.50 to 0.99 is acceptable in social science research especially when most of the explanatory variables are statistically significant. The only caveat to this is that the high R-squared should not be caused by spurious causation or multi-collinearity among the explanatory variables.

5. Conclusion

This paper examined the acceptable R-squared in empirical modelling in social science research with particular focus on whether a low R-squared is acceptable. Assuming the R-squared is the only decision rule being considered, the paper argued that a low R-squared of at least 0.10 is acceptable in social science empirical modelling provided that some or most of the explanatory variables are statistically significant. This means that regression models that have a low R-squared are good models if some of the explanatory variables are statistically significant. Therefore, a regression model in social science research should not be discarded solely because it has a low R-squared. The decision on whether the model is good or not should take into account the statistical significance of the explanatory variables in the model. It is hoped that this commentary article will guide young researchers who are beginners in social science empirical research. I wrote this piece specifically for them.

References

Cameron, A. C., & Windmeijer, F. A. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of econometrics*, *77*(2), 329-342.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623.

Cornell, J. A., & Berger, R. D. (1987). Factors that influence the value of the coefficient of determination in simple linear and nonlinear regression models. *Phytopathology*, *77*(1), 63-70.

CSCU (2005). Assessing the Fit of Regression Models. The Cornell Statistical Consulting Unit

Ferligoj, A., & Kramberger, A. (1995). Some Properties of R 2 in Ordinary Least Squares Regression.

Figueiredo Filho, D. B., Júnior, J. A. S., & Rocha, E. C. (2011). What is R2 all about?. *Leviathan (São Paulo)*, (3), 60-68.

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian regression models. *The American Statistician*.

Gujarati, D. N., Porter, D. C., & Gunasekar, S. (2012). *Basic econometrics*. Tata mcgraw-hill education.

Hagquist, C., & Stenbeck, M. (1998). Goodness of fit in regression analysis–R 2 and G 2 reconsidered. *Quality and Quantity*, *32*(3), 229-245.

Hagle, T. M., & Mitchell, G. E. (1992). Goodness-of-fit measures for probit and logit. *American Journal of Political Science*, 762-784.

Hill, R. C., Griffiths, W. E., & Lim, G. C. (2018). *Principles of econometrics*. John Wiley & Sons.

Ijomah, M. A. (2019). On the Misconception of R-square for R-square in a Regression Model. International Journal of Research and Scientific Innovation (IJRSI), 6(12), 71-76.

King, G. (1986). How not to lie with statistics: Avoiding common mistakes in quantitative political science. *American Journal of Political Science*, 666-687.

King, G. (1990). Stochastic Variation: A Comment on Lewis-Beck and Skalaban's "The R-Squared". *Political Analysis*, *2*, 185-200.

Lewis-Beck, M. S., & Skalaban, A. (1990). The R-squared: Some straight talk. *Political Analysis*, *2*, 153-171.

Lewis-Beck, C., & Lewis-Beck, M. (2015). *Applied regression: An introduction* (Vol. 22). Sage publications.

Maydeu-Olivares, A., & Garcia-Forero, C. (2010). Goodness-of-fit testing. *International encyclopedia of education*, *7*(1), 190-196.

Miles, J. (2005). R-squared, adjusted R-squared. *Encyclopedia of statistics in behavioral science*.

Onyutha, C. (2020). From R-squared to coefficient of model accuracy for assessing" goodness-of-fits". *Geoscientific Model Development Discussions*, 1-25.

Reisinger, H. (1997). The impact of research designs on R2 in linear regression models: an exploratory meta-analysis. *Journal of Empirical Generalisations in Marketing Science*, *2*(1).

Spiess, A. N., & Neumeyer, N. (2010). An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. *BMC pharmacology*, *10*(1), 1-11.