A short comment on the use of R^2_{adj} in Social Science

A short comment on the use of R^2_{adj} in Social Science

Un breve comentario sobre el uso de R^2_{adj} en Ciencias Sociales

ABSTRACT

It is a common practice to prefer R^2_{adj}, over R^2 to assess the explainability power of a statistical regression model among social scientists, especially for one having more than one independent variables. However, this preference is not advantageous at all times because the usage of R^2_{adj} may end up in negative coefficients making them non-interpretable. A Monte Carlo simulation experiment is used to appraise the behavior of these adjusted versions of R^2 for different numbers of independent variables. It has been found that almost all of the selected adjusted version of R^2 produces negative coefficients.

KEYWORDS: R^2, R^2_{adj}, statistics, models, social sciences.

RESUMEN

Es una práctica común preferir R^2_{adj}, sobre R^2 para evaluar el poder de explicación de un modelo de regresión estadística entre los científicos sociales, especialmente para una que tiene más de una variable independiente. Sin embargo, esta preferencia no es ventajosa en todo momento porque el uso de R^2_{adj} puede terminar en coeficientes negativos, lo que los hace no interpretables. Se utiliza un experimento de simulación de Monte Carlo para evaluar el comportamiento de estas versiones ajustadas de R^2 para diferentes números de variables independientes. Se ha encontrado que casi toda la versión ajustada seleccionada de R^2 produce coeficientes negativos.

PALABRAS CLAVE: R^2, R^2_{adj}, estadísticas, modelos, ciencias sociales.

Ahmed F. Siddiqi
University of Central Punjab, Lahore, Pakistan
dr.ahmedsiddiqi@ucp.edu.pk

Muhammad Salman Shabbir
Universiti Sains Malaysia, Penang, Malaysia
salman@usm.my

Faiza Khalid
University of Central Punjab, Lahore Pakistan
salman@usm.my

Rabia Salman
University Utara Malaysia
aarabiaa@gmail.com

Muhammad Farooq Shabbir
Dalian University of Technology, Liaoning, China.
farooq.shabbir@iub.edu.pk

ARTÍCULO RECIBIDO: 14 DE MARZO DE 2019
ARTÍCULO ACEPTADO PARA PUBLICACIÓN: 20 DE MARZO DE 2019
ARTÍCULO PUBLICADO: 30 DE ABRIL DE 2019

Introduction

The R^2 statistic, when used in a regression or ANOVA context, is appealing because it summarizes how well the regression model explains the data in an easy-to-understand way. R^2 statistics are also useful to gauge the effect of changing a model. However, being an increasing function of k (the number of independent variables in the regression model), its usage to assess the explanatory power of a newly added independent variable is highly criticized. There has been a tendency to use its adjusted version, R^2_{adj}, which is duly adjusted for the number of independent variables in the model, to make it usable to assess the explanatory power of any added independent variable. The adjusted version, quite popular in its usage is marred, however, with illogical results, including negative values. The current article is an attempt to highlight this and related issues while using the adjusted version of R^2.

Background

An omnipresent feature of any regression analysis is to calculate and interpret its R^2 to assess its efficacy, reliability and applicability. For a typical linear additive regression model to measure a variable y, with k where $(k \geq 1)$ and given independent variables $x_1, x_2, \ldots, x_k$, the mean value of y in a sample of size n where $(n \geq k)$, in standard symbols, is given by

$$E(y) = \hat{\beta}_0 + x_1\hat{\beta}_1 + x_2\hat{\beta}_2 + \ldots + x_k\hat{\beta}_k \qquad (1)$$

where $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are the estimated values for the unknown parameters of the regression model, giving the partial dependencies of y upon these k independent variables with $k \leq n$.

The variation in y, from its mean, given by $\sum_i^n (y_i - \overline{y})^2$, which is believed to be due to the given k independent variables (if the choice of $x_1, x_2, \ldots, x_k$ is true), may be decomposed into variation duly *explained* by the model, i.e. $\sum_i^n (\hat{y}_i - \overline{y})^2$, and the variation $\sum_i^n (y_i - \hat{y}_i)^2$ could not be explained by the model. Details of this decomposition and derivation is out of the scope of current writing and is available in any standard Statistics textbook on regression models, like Draper & Smith (2014); Richard B. Darlington & Hayes (2017); Rincon-Flores et al., (2018).

The ratio of this explained variation to the total variation, may be used to assess the proportion of variation in y, explained by the given independent variable, $x_1, x_2, \ldots, x_k$, and is given by

$$\frac{\sum_i^n (\hat{y}_i - \overline{y})^2}{\sum_i^n (y_i - \overline{y}_i)^2} = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \overline{y}_i)^2} \qquad (2)$$

This is commonly known as R^2 in statistical literature. Obviously, more the explanation, more is the value of R^2 and vice versa. In other words, more the value of R^2, more expletive the model is considered to be, in explaining the variation in y. Similarly, one may relate the value of R^2 with the goodness of choice of k independent variables in the regression model. In that case, more the value of R^2 better is the researchers' choice of independent variables and vice versa. As a matter of fact, these are the interpretations which make the calculation of R^2 an omnipresent feature of any regression analysis.

Intuitively, R^2 is related to the explanatory power of the estimated form of the regression model in Eq. (1). As a matter of fact, R^2 is a function of; (i) shape of the model as being linear or non-linear, (ii) construction of the model i.e. how the constituent independent variables are combined together, and (iii) number of independent variables.

Apart from the theoretical universality of the concept of R^2, there have been a few exceptions where the calculation of R^2 did not produce meaningful results. For example, Kvålseth (1985) wrote a cautionary note on the use of R^2 for an intercept-less regression model, especially when it is to be compared with intercept-present regression models. Nagelkerke (1991); Cox & Snell (1989) established that R^2 does not achieve its maxima for binary data and developed modified versions for such data types. Similarly, McCullagh (1980); Cameron & Windmeijer (1996) modified it for ordinal data sets. Helland (1987); Snyder & Lawson (1993); Carter (1979); Fan (2001); Thompson (1999); Yin & Fan (2001) established that R^2 always has a positive bias as an estimator of ρ^2, the squared correlation coefficient in population, and this bias is corrected by using the adjusted version of the R^2. Mittlböck & Waldhör (2000) made some adjustments to apply R^2 in Poisson regression models. Montomery & Morrison (1973) showed that R^2 is a positively biased estimator of the true R^2 and additionally, it does not penalize the likelihood function for having additional variables. Yin & Fan (2001) commented on the cross-validational power of R^2 which, as per their observations, drops when it is applied to an independent sample other than from which it is obtained. Glass & Hopkins (1996) and Pituch & Stevens (2015) observed this drop and refered it to statistical bias. So, the additions of independent variables to a model often increase the value of R^2 even when the additional variables have no explanatory power (Cameron & Windmeijer, 1996). This makes the R^2 a weak assessor of the efficacy and suitability of adding any new variable. An adjusted version of R^2 is developed to address this shortcoming. Similarly, Mittlböck & Waldhör (2000) observed that for a small sample size, relative to the number of independent variables, R^2 may be seriously inflated and may need to be adjusted according to the number of independent variables in the model.

To remove, at least statistically, the biases associated with R^2, various adjustments have been proposed. Following is a list, adopted mainly from Snyder & Lawson (1993);Yin & Fan (2001) and Leach & Henson (2007), showing a few such adjustments. Academic literature splits these adjustments into formulas used to estimate ρ^2; like Smith, Ezekiel, Wherry-1 & 2, Pratt, Olkin & Pratt and Claudy are developed as the sample estimates of ρ^2. While rest of the formulae are developed to adjust R^2 to increase predictive power for cross validation purposes.

Table 1. Different Adjustments made in R^2

Called as	Cited in	Expression
Smith	Ezekiel (1929)	$1-\frac{n}{n-k}(1-R^2)$
Ezekiel	Ezekiel (1929)	$1-\frac{n}{n-k-1}(1-R^2)$
Wherry-1	Ayabe (1985);J. Stevens (1996);Pituch & Stevens (2015)	$1-\frac{n}{n-k-1}(1-R^2)$
Wherry-2	Wherry (1931)	$1-\frac{n-1}{n-k}(1-R^2)$
Pratt	Claudy (1978)	$1-\frac{(n-3)(1-R^2)}{n-k-1}\left(1+\frac{2(1-R^2)}{n-k-2.3}\right)$
Olkin-Pratt	Olkin, Pratt, & others (1958)	$R^2-\frac{k-2}{n-k-1}(1-R^2)-\frac{2(n-3)}{(n-k-1)(n-k+1)}(1-R^2)^2$
Claudy-1	Claudy (1978)	$(2\rho-R)^2$
Claudy-2	Claudy (1978)	$1-\left(\frac{n-1}{n-k-1}\right)\left(\frac{n-2}{n-k-2}\right)\left(\frac{n-1}{n}\right)(1-R^2$
Claudy-3	Claudy (1978)	$1-\frac{(n-4)(1-R^2)}{n-k-1}\left(1+\frac{2(1-R^2)}{n-k+1}\right)$
Lord-1	Newman & others (1979);Uhl & Eisenberg (1970)	$1-\frac{n+k+1}{n-k-1}(1-R^2)$
Lord-2	Newman & others (1979);Kennedy (1988)	$1-\frac{n+k+1}{n-k-1}(1-R^2)\frac{n-1}{n}$
Darlington-Stein	Richard B Darlington (1968);Stein (1960)	$1-\left(\frac{n-1}{n-k-1}\right)\left(\frac{n-2}{n-k-2}\right)\left(\frac{n+1}{n}\right)(1-R^2$
Rozeboom-1	Rozeboom (1978)	$1-\left(\frac{n+k}{n-k}\right)(1-R^2$
Rozeboom-2	Rozeboom (1981)	$\rho^2\left(1+\left(\frac{k}{n-k-2}\right)\left(\frac{1-\rho^2}{\rho^2}\right)\right)^{-1}$
Browne	Schmitt (1982);Yin & Fan (2001)	$\frac{(n-k-3)\rho^4+\rho^2}{(n-2k-2)\rho^2+\rho}$

The expression proposed as Wherry-1, or Ezekiel is being implemented by popular statistical packages for computing R^2_{adj} in multiple regression procedures.

There appears to be a lack of consensus in the literature on which method is most appropriate under what circumstances for adjusting the R^2 (Yin & Fan, 2001). For example, Kromrey & Hines (1996) suggested that Browne formula is superior; Huberty & Mourad (1980) equated the performance of formula propounded by Olkin and Olkin & Pratt.

The adjustments do make the R^2 oblivious of the number of independent variables in the model, however, these adjustments invite other issues. For example, Ayabe (1985), Kennedy (1988) and

REVISTA SAN GREGORIO, 2019, No.30,SPECIAL EDITION.MARCH (24-31), ISSN: 1390-7247; eISSN: 2528-7907

J. Stevens (1996) observed that the Ezekiel formula was wrongly refereed as the Wherry's. Then researchers used these formulas without distinguishing them as estimators of ρ^2 or future sample effect estimator (for cross validation purposes). Barten (1962) showed that R^2_{adj} is also positively biased. Further, there are some combinations of n and k for which these adjusted versions of R^2 produce negative coefficients, which make them logically implausible and non-Interpretable. It is attempted here in this paper to appraise the performance of these adjusted versions of R^2 for their logical plausibility.

Methodology

As Yin & Fan (2001) lament, one of the major limitations in appraising these coefficients is the non-availability of real data which satisfies requisite statistical conditions of data-suitability, e.g. normality, multicollinearity, homoskedasticity, auto-correlation, etc., and is free of confounding factors which researchers can not control.

The present study is based upon a Monte Carlo simulation experiment to generate data which at one hand is based upon normal probability distribution, and different statistical conditions are satisfied, both of which are required to develop *good* regression models, as described above. As a matter of fact, the normality is assured by using Jarque & Bera (1987) test (as per the findings of Siddiqi (2014);Jarque (2011)), the multicollinearity is assured by not letting any variable to enter in the simulation experiment whose variance-inflation-factor (VIF) is greater than 10 (as per the suggestions of Gujarati (2009)), the homoskedasticity is assured by using robust standard errors (as per the suggestions of Long & Ervin (2000);Breusch & Pagan (1979)), the auto-correlation is controlled by not letting any data set enter into the experiment which produce a Durbin & Watson (1950) test value beyond the range of (1.8,2.2) (as per the suggestion of Gujarati (2011)). All these filtration and constraints are applied to make the comparison fair and based upon data which is not faulty in itself. *R*, version 3.5.1, with its libraries like *normtest*, *lmtest*, *car*, *sandwich* besides its default, are used for conducting this experiment.

All the coefficients, listed above, are calculated for a sample of size 100, a large sample, and for each value of k, ranging from 1 to 80. The results are shown in the figure below

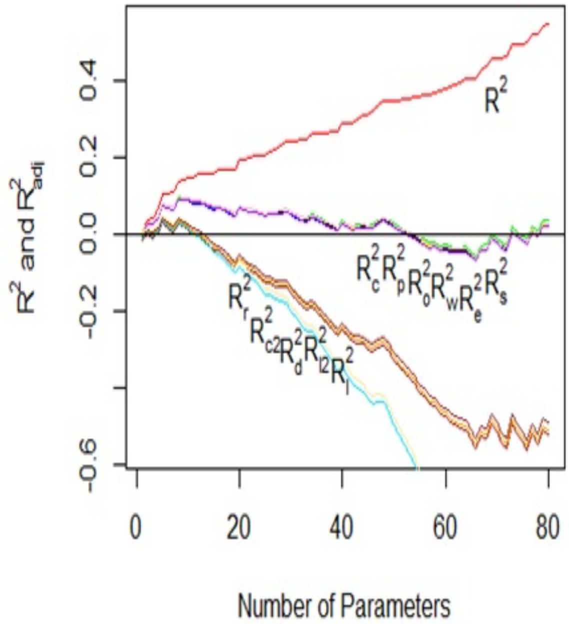

Figure 1. Monte Carlo Simulation Based Comparison of R^2 and R^2_{adj}

The results of this Monte Carlo simulation experiment are shown in a two dimensional graph (shown in the figure); the horizontal axis measures k, in a range from 1 to 80, while the vertical axis measures R^2 and R^2_{adj}. For each different version of R^2 as discussed in the table, a seperate line is drawn. There are few things which are obvious from this figure:

1. R^2 is an increasing function of k, the number of independent variables in the regression model. In other words, for every increase in number of independent variables, there is a definite increase in R^2. As a matter of fact, this is the biggest criticism on R^2. And this is the reason researchers avoid using it.
2. None of the values of R^2 fall below the zero level line, i.e. all values of R^2 are positive. .
3. R^2_{adj} is not an increasing function in number of parameters. In other words, the variations in R^2_{adj} are proportional to the importance of the corresponding independent variables i.e. more important the independent variable, more is the change in R^2_{adj} and vice versa. So, this may be a good indicator of assessing the addition in the suitability of the regression model by adding any new independent variable.
4. There exist values of R^2_{adj} which fall below the zero level line, which indicates the negative values for R^2_{adj}. This seems quite un-natural and illogical.

Concluding Remarks

The use of $R^2$is omni present for assessing and establishing the appropriateness and the strength of the independent variables used in a regression model. However, its usage is usually made in some adjusted versions which have been derived for its incapacity to assess the appropriateness of an additional independent variable, where it has been seen that it always increases for an addition of independent variable(s) in the regression model. As a matter of fact, the contention is correct as the R^2 has dully been established as an increasing function of the independent variables, so it cannot capture the real increase in the appropriateness of the regression model upon an increase in the number of independent variables.

To address this shortcoming, many adjustments have been presented to make it more representative of the appropriateness of the regression model, especially for making decisions about the inclusion, or exclusions, of independent variables. Interestingly, these adjustments, do adjust $R^2$for this shortcoming but instill other issue(s). The current article attempts to highlight one such problem where they produce negative coefficient for certain values of *n* and *k*.

A Monte Carlo simulation experiment is conducted to generate data which abide by all the requisite assumptions for a typical regression model, e.g. normality, multicollinearity, auto-correlation, homoskedasticity, etc. Once the data is generated, coefficients for these adjusted versions are calculated, and so is R^2, where n = 100 and k ranges from 1 to 80. To visualize the dynamics of these R^2_{adj}, for their comparison, a two-dimensional graph (Figure 1) with k along horizontal axis and their corresponding coefficients along vertical axis is used.

The graph explicitly shows the increasing nature of R^2 and the negative coefficients for almost all of its the adjusted versions. A negative value for a squared entity is illogical and non-interpretable. And, this is what makes their usage, at least mathematically, "objectionable".

Such results prohibit the rampant usage of these adjusted versions of R^2. This article only speculates upon the observation that usage of R^2_{adj} gives negative values. It should be in no way, taken as a relative comparison between the values obtained by R^2 and R^2_{adj}, it only highlights the fact that it gives negative values (which is considered mathematically illogical), hence it is not recommendable. Secondly this study is based on a larger sample i.e. n=100. The results may vary when smaller size sample is used. R^2 value happens to be 0.4 but result may be different for regression models having R^2 closer to 1.

References

Ayabe, C. R. (1985). Multicrossvalidation and the jackknife in the estimation of shrinkage of the multiple coefficient of correlation. *Educational and Psychological Measurement*, *45*(3), 445–451.

Barten, A. (1962). Note on unbiased estimation of the squared multiple correlation coefficient. *Statistica Neerlandica*, *16*(2), 151–164.

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, 47(5), 1287–1294.

Cameron, A. C., & Windmeijer, F. A. G. (1996). R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics*, *14*(2), 209–220.

Carter, D. S. (1979). Comparison of different shrinkage formulas in estimating population multiple correlation coefficients. *Educational and Psychological Measurement*, *39*(2), 261–266.

Claudy, J. G. (1978). Multiple regression and validity estimation in one sample. *Applied Psychological Measurement*, *2*(4), 595–607.

Cox, D., & Snell, E. (1989). *The analysis of binary data* (2nd Edition.). Chapman; Hall, London.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, *69*(3), 161-182.

Darlington, R. B., & Hayes, A. F. (2017). *Regression analysis and linear models; concepts, applications and implementations*. The Guilford Press, New York.

Rincon-Flores, E. G., Gallardo, K., & Fuente, J. M. D. L. (2018). Strengthening an Educational Innovation Strategy: Processes to Improve Gamification in Calculus Course through Performance Assessment and Meta-evaluation. *International Electronic Journal of Mathematics Education, 13*(1), 1-11.

Draper, N. R., & Smith, H. (2014). *Applied regression analysis* (Vol. 326). John Wiley & Sons, New York.

Durbin, J., & Watson, G. S. (1950). Testing for serial correlation in least squares regression: I. *Biometrika*, *37*(3/4), 409–428.

Ezekiel, M. (1929). The application of the theory of error to multiple and curvilinear correlation. *Journal of the American Statistical Association*, *24*(165A), 99–104.

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *The Journal of Educational Research*, *94*(5), 275–282.

Glass, G. V., & Hopkins, K. D. (1996). Statistical methods in education and psychology. Needham Heights, MA, US: Allyn & Bacon.

Gujarati, D. N. (2009). *Basic econometrics*. Tata McGraw-Hill Education, New Delhi.

Gujarati, D. N. (2011). Econometrics by example, Palgrave MacMilillan, New York.

Helland, I. (1987). On the interpretation and use of R^2 in regression analysis. *Biometrics*, *43*(1), 61–69.

Huberty, C. J., & Mourad, S. A. (1980). Estimation in multiple correlation/prediction. *Educational and Psychological Measurement*, *40*(1), 101–112.

Jarque, C. M. (2011). Jarque-bera test. In *International encyclopedia of statistical science* (pp. 701–702). Springer, New York.

Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique*, 55(2), 163–172.

Kennedy, E. (1988). Estimation of the squared cross-validity coefficient in the context of best subset regression. *Applied Psychological Measurement*, *12*(3), 231–237.

Kromrey, J. D., & Hines, C. V. (1996). Estimating the coefficient of cross-validity in multiple regression: A comparison of analytical and empirical methods. *The Journal of Experimental Education*, *64*(3), 240–266.

Kvålseth, T. O. (1985). Cautionary note about R^2. *The American Statistician*, *39*(4), 279–285.

Leach, L. F., & Henson, R. K. (2007). The use and impact of adjusted r2 effects in published regression research. *Multiple Linear Regression Viewpoints*, *33*(1), 1–11.

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*(3), 217–224.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Ser. B*, *42*(2), 109–142.

Mittlböck, M., & Waldhör, T. (2000). Adjustments for r2-measures for poisson regression models. *Computational Statistics & Data Analysis*, *34*(4), 461–472.

Montomery, D., & Morrison, D. (1973). A note on adjusting R^2. *The Journal of Finance*, *28*(4), 1009–1013.

Nagelkerke, N. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692.

Newman, I., & others. (1979). A monte carlo evaluation of estimated parameters of five shrinkage estimate formuli. *Multiple Linear Regression Viewpoints*, *9*(5), 57–74.

Olkin, I., Pratt, J. W., & others. (1958). Unbiased estimation of certain correlation coefficients. *The Annals of Mathematical Statistics*, *29*(1), 201–211.

Pituch, K. A., & Stevens, J. P. (2015). *Applied multivariate statistics for the social sciences: Analyses with sas and ibm's spss*. Routledge.

Rozeboom, W. W. (1978). Estimation of cross-validated multiple correlation: A clarification. *Psychological Bulletin*, *85*(6), 1348.

Rozeboom, W. W. (1981). The cross-validational accuracy of sample regressions. *Journal of Educational Statistics*, *6*(2), 179–198.

Schmitt, N. (1982). Formula estimation of cross-validated multiple correlation, Presented at the Annual Meeting of the American Psychological Association, Washington, DC..

Siddiqi, A. (2014). An observatory note on tests for normality assumptions. *Journal of Modelling in Management*, *9*(3), 290–305.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *The Journal of Experimental Education*, *61*(4), 334–349.

Stein, C. (1960). Multiple regression. in Olkin et al.(Eds.), Contributions to Probability and Statistics. Stanford: Stanford University Press.

Stevens, J. (1996). Applied multivariate analysis for the social sciences, Mahwa, New Jeresey.

Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology*, *9*(2), 165–181.

Uhl, N., & Eisenberg, T. (1970). Predicting shrinkage in the multiple correlation coefficient. *Educational and Psychological Measurement*, *30*(2), 487–489.

Wherry, R. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *The Annals of Mathematical Statistics*, *2*(4), 440–457.

Yin, P., & Fan, X. (2001). Estimating r 2 shrinkage in multiple regression: A comparison of different analytical methods. *The Journal of Experimental Education*, *69*(2), 203–224.