A New (Old) Goodness of Fit Metric for Multidimensional Outcomes

Tommy Jones[*] Mark Meyer[†]

2021-01-31

1 Introduction

The coefficient of determination—R^2—is the most popular goodness of fit metric for linear models. It has appealing properties such as a lower bound of zero, an upper bound of one, and interpretable as the proportion of varience in the outcome accounted for by the model. R^2's appeal is such that nearly all statistical software reports R^2 by default when fitting linear models.

The standard, and model free, definition of R^2 is

$$R^2 \equiv 1 - \frac{SS_{resid.}}{SS_{tot.}} \tag{1}$$

or equivalently

$$R^2 \equiv 1 - \frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{2}$$

In (1) $SS_{resid.}$ is called the *residual sum of squares* and $SS_{tot.}$ is called the *total sum of squares.* And in (2) $\bar{y}$ is the sample mean and $\hat{y}_i$ is the value predicted for observation y_i.

The definitions in In (1) and (2) are "model free". But for a linear model, they are alegebraically equivalent to

[*] Dept. of Computational and Data Sciences, George Mason University

[†] Dept. of Mathematics and Statistics, Georgetown University

$$R^2 = \frac{V_{i=1}^{N}\hat{y}_i}{V_{i=1}^{N}y_i} \tag{3}$$

where $V_{i=1}^{N}z_i = \frac{1}{N-1}\sum_{i=1}^{N}(z_i - \bar{z})^2$. Equation (3) is the proportion of variance in y accounted for by the (linear) model. Researchers have used and extended (3) to develop R^2 metrics for a range of models such as logistic regression (Hu, Palta, and Shao 2006), mixed models (Piepho 2019), Bayesian regression models (Gelman et al. 2018), and more.

To our knowledge our research is the first time anyone has proposed a variation of R^2 for models predicting an outcome in multiple dimensions—where each $\boldsymbol{y}_i$ is a vector and $\boldsymbol{Y}$ is a matrix. Multidimensional outcomes occur in settings such as [need good examples here... simultaneous equations? neural networks? topic models?] Our R^2 relies on a geometric interpretation of (1). As a result, it is also model free.

2 A Geometric Interpretation of R^2

The numerator and denominator in (2) may be viewed as sums of squared Euclidean distances in $\mathbb{R}_1$. Specifically, $SS_{tot.}$ is the total squared-Euclidean distance from each y_i to the mean outcome, $\bar{y}$. Then $SS_{resid.}$ is the total squared-Euclidean distance from each y_i to its predicted value under the model, $\hat{y}_i$.

As a reminder, for any two points $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}_M$, the Euclidean distance between $\boldsymbol{p}$ and $\boldsymbol{q}$ is

$$d(\boldsymbol{p}, \boldsymbol{q}) = \sqrt{\sum_{i=1}^{M}(p_i - q_i)^2} \tag{4}$$

Using the notation from (4), we can rewrite (1) and (2) as

$$R^2 = 1 - \frac{\sum_{i=1}^{N} d(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i)^2}{\sum_{i=1}^{N} d(\boldsymbol{y}_i, \bar{\boldsymbol{y}})^2} \tag{5}$$

Equation (5) extends R^2 to cover outcomes in $\mathbb{R}_M$ while preserving the calculation from (2) for outcomes in $\mathbb{R}_1$.

Fig. 1 visualizes the geometric interpretation of R^2 for outcomes in $\mathbb{R}_2$. The left image represents $SS_{tot.}$: the red dots are data points ($\boldsymbol{y}_i$); the black dot is the vector of means ($\bar{\boldsymbol{y}}$); the line segments represent the Euclidean distance from each $\boldsymbol{y}_i$ to $\bar{\boldsymbol{y}}$. $SS_{tot.}$ is obtained by squaring the length of each line segment and then adding the squared segments together. The right image represents $SS_{resid.}$: the blue dots are the fitted values ($\hat{\boldsymbol{y}}_i$); the line segments represent the Euclidean distance from each $\hat{\boldsymbol{y}}_i$ to its corresponding $\boldsymbol{y}_i$. $SS_{resid.}$ is obtained by squaring the length of each line segment and then adding the squared segments together.

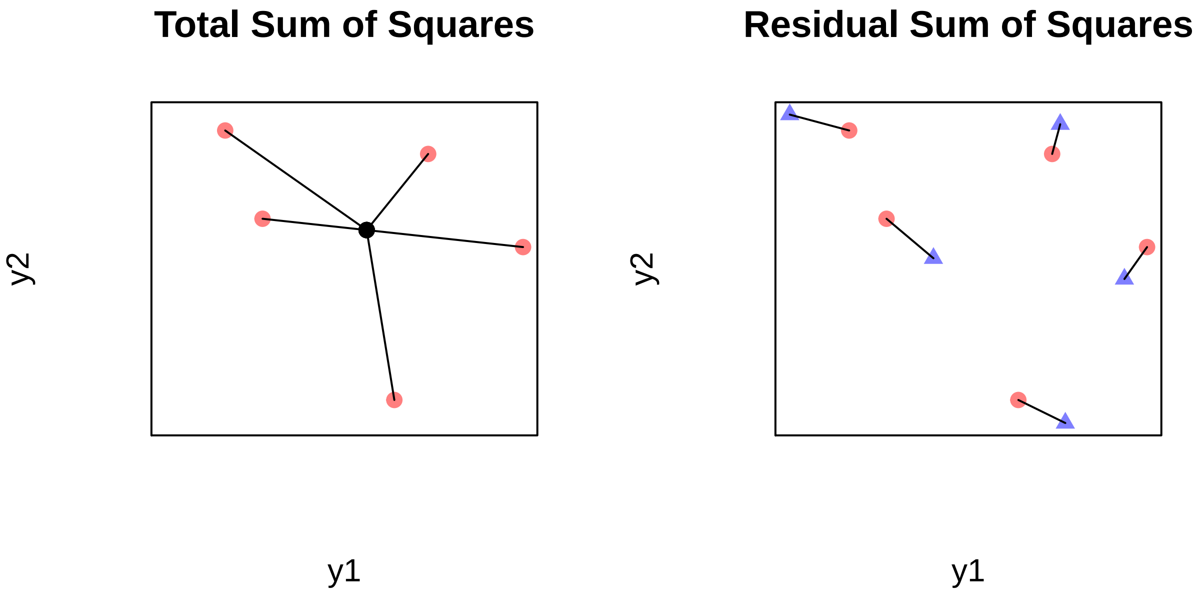

Figure 1: Visualizing the geometric interpretation of R-squared. Sum up the squared length of each line segment for the total (left) or residual (right) sums of squares. This figure corresponds to an R-squared of 0.87

3 Properties

You can view the R^2 in (5) as the proportion of total squared distance from each observation to its mean that is accounted for by the model, similar to the "proportion of explained variance" interpretation.

This R^2 is maximized at 1—representing perfect predictions. Unlike traditional R^2, it does not have a lower bound, as can be the case in certain circumstances with tradtional R^2 (Barten 1987) (Cameron and Windmeijer 1997). Losing the lower bound of zero does not negatively impact interpretation. When $R^2 = 0$, then $SS_{resid.} = SS_{tot.}$; the model is no better than guessing the mean outcome. So negative values of R^2 mean the model is *worse* than guessing the mean outcome for every observation. Its interpretation remains straightforward.

The R^2 given in (5) is sensitive to the scale of the dimensions of $\boldsymbol{Y}$. This is a common property of Euclidean distance, exacerbated by squaring. Possible mitigating steps may be standardizing the columns of $\boldsymbol{Y}$ before modeling or standardinzing both the columns of $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ after modeling. When one standardizes may lead to different results, an exploration of which is out of scope for this note.

4 Discussion

Viewing R^2 as a ratio of distances motivates a new direction for extended definitions of R^2. Researchers may wish to explore R^2 calculations based other distance measures more appropriate for other settings. For example, if one's model estimates probabilities perhaps Hellinger distance (Hellinger 1909) would be more appropriate. If scale between outcome variables is of concern, maybe an R^2 leveraging Mahalanobis distance (Mahalanobis 1936) has advantage over variable normalization as discussed above.

We have shown that R^2 can be calculated for multivariate outcomes from a geometric interpretation of (1). Since we use the standard definition of R^2, this approach does not alter existing issues motivating alternate definitions of R^2 for, e.g., Bayesian models or nonlinear models. It is worth noting that scale differences in outcome variables can swamp the calculation. We leave an exploration of remediation strategies to future work. We have implemented the R^2 metric in (5) in a package for the R programming language called `mvrsquared` (Jones 2020).

References

Barten, Anton P. 1987. "The Coeffecient of Determination for Regression without a Constant Term." In *The Practice of Econometrics*, edited by Heijmans R. and Neudecker H., 15:187—189. International Studies in Economics and Econometrics. Springer, Dordrecht. https://doi.org/10.1007/978-94-009-3591-4/_12.

Cameron, A. Colin, and Frank A.G. Windmeijer. 1997. "An R-squared measure of goodness of fit for some common nonlinear regression models." *Journal of Econometrics* 77 (2): 329–42. https://doi.org/10.1016/s0304-4076(96)01818-0.

Gelman, Andrew, Ben Goodrich, Jonah Gabry, and Aki Vehtari. 2018. "R-squared for Bayesian Regression Models." *The American Statistician* 73 (3): 1–6. https://doi.org/10.1080/00031305.2018.1549100.

Hellinger, Ernst. 1909. "New Definition of the Theory of Quadratic Forms of Infinitely Many Different "a Changes." *Journal F "u R Pure and Applied Mathematics* 1909 (136). De Gruyter: 210–71.

Hu, Bo, Mari Palta, and Jun Shao. 2006. "Properties of R^2 statistics for logistic regression." *Statistics in Medicine* 25 (8): 1383–95. https://doi.org/10.1002/sim.2300.

Jones, Tommy. 2020. "Mvrsquared." 2020. https://CRAN.R-project.org/package=mvrsquared.

Mahalanobis, Prasanta Chandra. 1936. "On the Generalized Distance In Statistics." *Proceedings of the National Institute of Sciences of India* 2 (1): 49—55.

Piepho, Hans-Peter. 2019. "A coefficient of determination (R^2) for generalized linear mixed models." *Biometrical Journal* 61 (4): 860–72. https://doi.org/10.1002/bimj.201800270.