ISSN 1946-7664. MCFNS 2012
Vol. 4, Issue 2, pp. 73–76
Mathematical and Computational Forestry & Natural-Resource Sciences

AVAILABLE ONLINE AT HTTP://MCFNS.COM

Submitted: Jul. 13, 2012
Accepted: Sep. 26, 2012
Published: Sep. 30, 2012
Last Correction: Mar. 8, 2013

THE COMPARATIVE R^2 AND ITS APPLICATION TO SELF-REFERENCING MODELS

Mike Strub[1], Chris J. Cieszewski[2]

[1]*Adj. Professor,* [2]*Professor. Warnell School of Forestry & Natural Resources, University of Georgia, Athens, GA, USA*

Abstract. The traditional coefficient of determination or R^2 is the proportion of variation explained by a regression model versus the variation explained by the mean. This measure does not discriminate well between alternative self-referencing models such as site index curves. The traditional R^2 compares the variation explained by a model with the variation about the mean dependent variable, a very simple model. A generalized R^2 based on the proportion of the variation explained by the self-referencing model versus the variation explained by another simpler (yet more complicated than the mean) model provides better discrimination between candidate models. We call this generalized R^2 the Comparative Coefficient of Determination or Comparative R^2. Three growth series or plots from the South Africa Correlated Curve Trend Study are used to illustrate the difference between the traditional R^2 and the generalized R^2.

Keywords: Coefficient of determination, R-square, self-referencing models, goodness of fit, non-linear regression.

1 Introduction

The coefficient of Variation or R^2 is commonly used to evaluate the fit of regression models (Rao, 1973). R^2 is used to describe how well a mathematical model fits a dataset. For simple linear regression models it is the proportion of the variation in the data explained by that model beyond using only the mean of the data. As a proportion, R^2 must lie between zero and one with values close to zero indicating a poor fit to the data and numbers approaching one indicating an excellent fit. This measure of model fit has been commonly used to choose between alternative models.

This paper is motivated by the desire for a more realistic assessment of the fit of a special type of regression models, self-referencing models. One origin of self-referencing models lies in the modeling of site index curves (Bailey and Clutter, 1974; Cieszewski and Bailey, 2000). The basic idea is that an individual subject's (tree/plot/stand) height growth can be modeled by an equation having some parameters that are common (global) to all subjects, and other parameters that are subject specific (local). The local parameter varies by subject and assumes a unique value for each subject. Hence the total number of local parameters is equal to the number of subjects. One or more global parameters are common for all the subjects.

Because the self-referencing models offer good fit to the data and the R^2 is often close to one, the use of the usual R^2 to select the best model has limited value. The practitioner is often faced with the problem of choosing between two models, one with a slightly higher R^2 but more global parameters. The aim of this paper is to propose a modified R^2 that provides more powerful discrimination between self-referencing models. The generalized R^2 compares the self-referencing model or alternate model with a null hypothesis model.

Comparing the residual variation of a model with the variation of using only the average makes sense if an intercept is included in the model. However self-referencing models are often constrained to pass through zero height at zero age. For this reason it seems reasonable to compare self-referencing models with a straight line through zero, so the null model for R^2 involves estimating the slope of the line through zero rather than the horizontal line of mean of the observations. Since a separate curve is fit to each individual subject it also makes sense to allow the slope of the straight line through zero to vary by individual subject. The generalized R^2 in this paper will compare self-referencing model residual variation with residual variation from fitting a straight line through zero to each growth series. A second alternative model of interest is to use a single average model (often called Guide Curve) of the same form as the self-

referencing model. This comparison evaluates the value of the varying local parameter that takes on a unique value for each subject. We call this generalized R^2 the Comparative Coefficient of Determination or Comparative R^2 and abbreviate it as CR^2. The CR^2 can also be used to compare candidate model forms substituting the sloped line with a default model.

2 Data

The data for this study comes from the South African Correlated Curve Trend study of loblolly pine (Pinus Taeda). Only a small portion of the data was used to simplify analyses and to effectively illustrate the results on graphs. A single plot from each of three locations that was thinned prior to the onset of competition to 50 trees per hectare (135.5 trees per hectare) was used to model average height over time. The plots with lowest stocking were chosen to eliminate the impact of competition on height development. Figure 1 shows the relationship between height and age from seed for the three plot locations.

Figure 1: The relationship between average height and age at three low stocking locations of the Correlated Curve Trend Study.

This data is described in more detain in Strub and Bredencamp (1985).

3 Models

Null models used in this paper include the mean, a straight line through the origin for each growth series and the base model for the site index equation. The mean model is:

$$y_{ij} = \alpha + \varepsilon_{ij} \tag{1}$$

The straight line model is:

$$y_{ij} = \beta_i x_{ij} + \varepsilon_{ij} \tag{2}$$

The base model is:

$$y_{ij} = e^{\alpha - \frac{\beta}{x_{ij}}} + \varepsilon_{ij} \tag{3}$$

Where y is the response variable of interest (height), x is the covariate (age), α and β are model parameters to be estimated and e is the base of the natural logarithms. Also y_{ij} = the height at the j^{th} measurement of the i^{th} plot ($i = 1, \ldots, m, \quad j = 1, \ldots n_i$), x_{ij} is the age at the j^{th} measurement of the i^{th} plot and $\varepsilon_{ij} \sim N(0, \sigma^2)$. The base model was proposed by Schumacher (1939). Bailey and Clutter (1974) suggested two self-referencing forms of this model. The first site index equation form is anamorphic in shape with asymptotes that vary by growth series:

$$y_{ij} = e^{\alpha_i - \frac{\beta}{x_{ij}}} + \varepsilon_{ij} \tag{4}$$

Where α_i is a parameter specific to the i^{th} plot and β is a global parameter. The second form of the site index equation suggested by Bailey and Clutter (1974) is polymorphic in shape but with a single asymptote:

$$y_{ij} = e^{\alpha - \frac{\beta_i}{x_{ij}}} + \varepsilon_{ij} \tag{5}$$

Where α is the global parameter, $\beta_i$is a parameter that varies by plot, and all other variables previously defined. Cieszewski and Bailey (2000, eq. 14) suggested a polymorphic model with asymptotes that vary by growth series:

$$y_{ij} = e^{\alpha_i \left(1 - \frac{\beta}{x_{ij}}\right)} + \varepsilon_{ij} \tag{6}$$

All variables are as previously defined.

4 Generalization

R^2 is defined as the difference between the sum of squared errors about the mean and the sum of squared errors for a fitted regression model divided by the sum of squared errors about the mean.

$$R^2 = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 - \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_{ij})^2}{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2} \tag{7}$$

$$= 1 - \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_{ij})^2}{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2} \tag{8}$$

Where $\hat{y}_{ij}$ is the prediction of the j^{th} height for the i^{th} growth series and $\bar{y}$ is the overall mean.

$$\bar{y} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{m} n_i} \quad (9)$$

R^2 can be generalized into the CR^2 by considering two different models and the resulting estimates.

$$CR^2 = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij}-\hat{y}^*_{ij}\right)^2 - \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij}-\hat{y}_{ij}\right)^2}{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij}-\hat{y}^*_{ij}\right)^2} \quad (10)$$

$$= 1 - \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij}-\hat{y}_{ij}\right)^2}{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij}-\hat{y}^*_{ij}\right)^2} \quad (11)$$

Where $\hat{y}^*_{ij}$is the predicted height from the null model (the mean for the original definition of R^2) and $\hat{y}_{ij}$ is the predicted height for the alternative model.

5 Results

The six models were fitted to the Correlated Curve Trend Study data. Figure 2 shows the results for the three null models. Model (1) or the average height regardless of age used in the original definition of R^2 does not fit the data well at all and has high error sum of squares. This explains why R^2 for site index equations are usually quite high. Models (2) and (3) fit the data better resulting in smaller error sum of squares and a more discriminating CR^2.

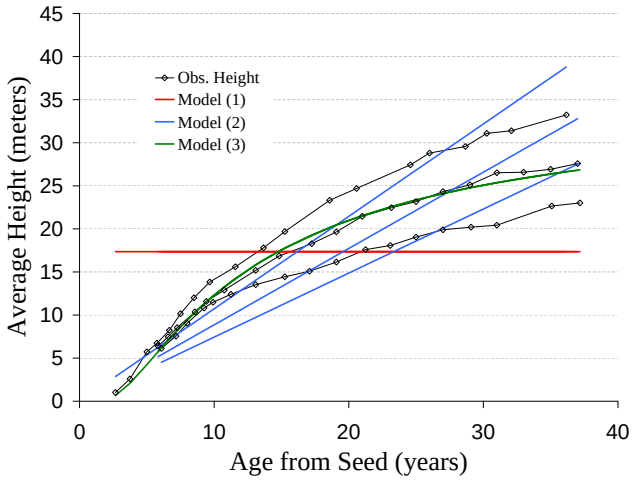

Figure 2: The relationship between average height and age at three low stocking locations of the Correlated Curve Trend Study.

Figure 3 shows the fit for the three site index equations, the anamorphic model (4), the polymorphic single asymptote model (5) and the polymorphic multiple asymptote model (6). Both models (4) and (6) fit all three plots well. Model (5) does not fit the top or bottom plot well at all.

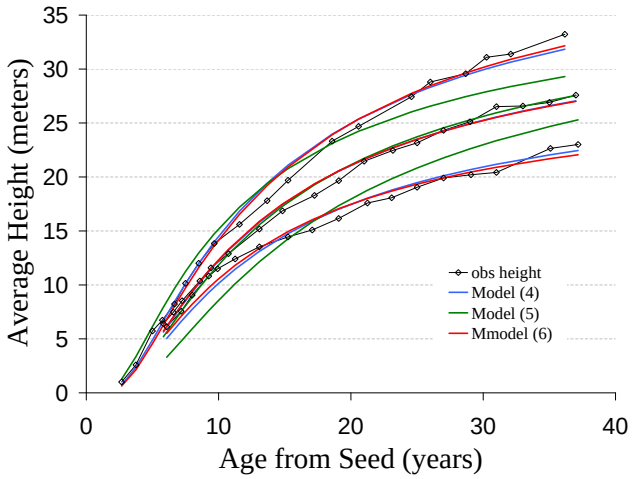

Figure 3: The relationship between average height and age at three low stocking locations of the Correlated Curve Trend Study.

Table 1 shows the usual R^2 with null model (1) in column 2, the generalized CR^2 for null model (2) in column 3 and the generalized CR^2 for null model (3) in column 4. Notice that the CR^2 provides greater discrimination between models, especially in the case of model (3) that does not fit the data well as shown in Figure 2.

Table 1: A comparison of the usual R2 and CR2 for three null models and three alternative models fit to the Correlated Curve Trend Study data.

	model (1)	model (2)	model (3)
model (4)	0.991	0.927	0.929
model (5)	0.956	0.629	0.639
model (6)	0.993	0.942	0.944

Another possibility is to consider model (5) the null model. Rows two and three of table 2 show the proportion of extra variation explained by the anamorphic model (4) and polymorphic multiple asymptote model (6) over the polymorphic single asymptote site index model (5). The last line of table 2 shows the proportion of additional variation explained by the polymorphic multiple asymptote model over the anamorphic model.

The anamorphic model explains 80.4% more variation than the polymorphic single asymptote model. The polymorphic multiple asymptote model explains 84.4% more variation than the polymorphic single asymptote model. The polymorphic multiple asymptote model explains 20.45% more variation than the anamorphic model.

Table 2: CR2 for the three site index models considered in this paper.

Comparison	CR2
model (4) versus model (5)	0.804
model (6) versus model (5)	0.844
model (6) versus model (4)	0.205

6 Discussion

The traditional R^2 indicates that all three models provide a good fit to the data, explaining over 95% of the error in the data. Examination of Figure 2 reveals that model (6) does not accurately fit the data, with too large a spread between plots at younger ages and too narrow a spread between plots at older ages. The lower 64% of variation explained when compared to straight lines through zero at time zero of the generalized R^2 is a better assessment of model fit to the data. Examination of Figures 2 indicates that both models (5) and (6) fit the data reasonably well., although the fit is not ideal for either model. The traditional R^2 seems to indicate a near perfect fit for both models with over 99% of the variation explained in both cases. The generalized R^2 assessment of 93% for model (4) and 94% for model (6) seems more reasonable. Comparison of the three candidate models with model (3) in table 1 provides further evidence of the poor fit of model (5). The anamorphic model (4) explains 93% more variation than a single Schumacher curve. The polymorphic multiple asymptote model (6) explains 94% more variation than a single Schumacher curve. The polymorphic single asymptote model (5) explains only 63% more variation than a single curve. Finally comparison of the three models in table 2 reinforces the poor fit of model (5) with model (4) explaining 80% more variation and model (5) explaining 84% more variation. Model (6) explains 20% more variation than model (4).

7 Conclusion

The Comparative R^2 provides better discrimination between site index models than the traditional R^2. CR^2 with null model (1) could be used instead of R^2 for any model constrained to be zero at the origin.

Acknowledgements

Thanks are due to the anonymous reviewers and Dr. Kim Iles for their helpful feedback on the earlier draft of this paper.

References

Bailey, R.L., and J.L. Clutter. 1974. Base-age invariant polymorphic site curves. For. Sci. 20:155–159.

Cieszewski, C.J., and R.L. Bailey. 2000. Generalized algebraic difference approach: Theory based derivation of dynamic site equations with polymorphism and variable asymptotes. For. Sci. 46:116–126.

Rao, C. R. (1973). Linear Statistical Inference and its Applications, 2nd ed. New York: Wiley.

Schumacher, F.X. 1939. A new growth curve and its application to timber yield studies. J. For. 37:819–820.

Strub, M. R., and B. V. Bredenkamp. 1985. Carrying capacity and thinning response of Pinus taeda in the CCT experiments. South African Forestry Journal, June 1985, pp. 6-11.