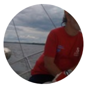
Evaluating the predictive power of regression models

ARTICLE *in* CANADIAN JOURNAL OF FISHERIES AND AQUATIC SCIENCES · MARCH 1996

Impact Factor: 2.29 · DOI: 10.1139/cjfas-53-3-490



Yves T. Prairie
Université du Québec à Montréal

Evaluating the predictive power of regression models[1]

Yves T. Prairie

Abstract: Regression models are routinely developed and used in aquatic sciences for predictive purposes. Although the traditional measures of predictive power for regression models (r^2, root mean square error) have well-defined statistical meanings, they do not necessarily provide an intuitive measure of the predictive utility of regression equations. It is proposed that an index of predictive power can be developed on the basis of the degree of categorical resolution a regression model can achieve. This index of resolution power is shown to increase nonlinearly with the familiar r^2 statistic, even under different distributional assumptions. This relationship also shows that the predictive power of models with $r^2 \leq 0.65$ is low and nearly constant but increases very rapidly for higher r^2 values, thereby justifying the search for additional explanatory variables even in models already explaining a large fraction of the variation.

Résumé : Les modèles de régression sont couramment développés et utilisés en sciences aquatiques à des fins prédictives. Bien que les mesures traditionnelles de puissance prédictive des modèles de régression (r^2, racine des moindres carrés moyen) aient des significations statistiques bien définies, elles ne fournissent pas nécessairement une mesure intuitive de l'utilité prédictive des équations de régression. Il est proposé qu'un indice de la puissance prédictive peut être développé sur la base du degré de résolution catégorique qu'un modèle de régression peut atteindre. Il est démontré que cet indice de puissance de résolution croît de façon non-linéaire avec le r^2, même sous différents postulats distributionnels. Cette relation démontre également que la puissance prédictive des modèles ayant un $r^2 \leq 0,65$ est faible et quasi constante mais croît très rapidement à des valeurs supérieures de r^2, justifiant ainsi la recherche de variables explicatives additionnelles même dans les modèles expliquant déjà une grande fraction de la variation.

The construction and use of regression models in aquatic sciences is common practice. Nearly 45% of the articles in the 1990 volume of the *Canadian Journal of Fisheries and Aquatic Sciences* contained at least one regression analysis. Although this statistical technique can be used for a variety of purposes (prediction, parameter estimation, identification of significant variables), it is nearly always useful to know the power of the regression model. Several measures of the strength of a relationship exist that are available from the computer output of most relevant software. Of those, the coefficient of determination (r^2) and the residual error variance (error mean square, σ_ε^2) are the most widely used. While the coefficients of correlation (r) and determination (r^2) provide relative measures of the degree of linear association between variables, the error mean square is an estimate of the absolute amount of uncertainty left. Despite the clear statistical basis and meaning of these measures, their interpretative meaning may be less obvious when cast in everyday terms. For example, how much better is a model explaining 90% of the variance ($r^2 = 0.9$) than one explaining only 80%? Is it only 10% better or is it twice as good? Similarly, is a model with half the error mean square of that of an alternative model really twice as powerful? Part of the answer to these questions clearly involves defining exactly what is meant by better and by powerful. The purpose of this note is to show that the utility of a regression model for prediction can be cast in simple terms with a measure called here the resolution power and that this measure may be closer to what we intuitively conceive predictive power to mean. I then present how this index of resolution power relates to the common r^2 statistic.

Suppose one wishes to know the predictive power of a simple linear regression model relating bacterial abundance to lake chlorophyll concentration (see Bird and Kalff 1984). One can address this question without referring to the r^2 value or the error mean square by asking "does this model allow me to predict on the basis of chlorophyll concentration that the bacterial abundance will be either low or high or does it allow me to predict that it will be low, intermediate, or high, or better still whether it will be very low, low, intermediate, high, or very high?" In other words, into how many classes (intervals) can the whole range of the dependent variable (here bacterial abundance) be divided so that predictions among each class are different from one another? Clearly, the larger the number of classes (i.e., finer resolution power), the greater the predictive utility of the model.

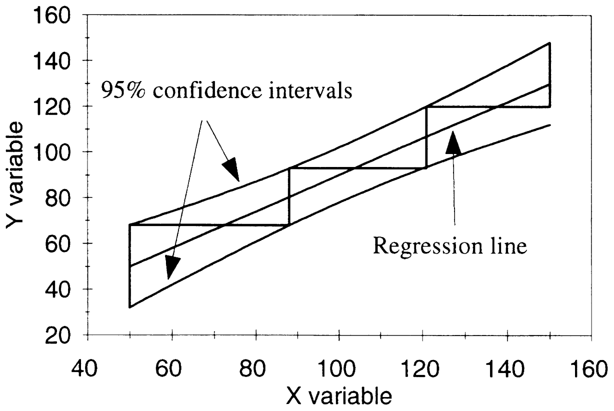
The degree of categorical resolution of a regression model will depend on its error variance relative to the total range covered by the dependent variable. If we define the width of a class interval as $2 \times t_{(\mathrm{df},p < 0.05)}$ multiplied by the standard error of estimate for an individual prediction (the EMS or σ_ε^2), then the number of distinguishable classes a regression model can discriminate is simply the ratio of the total range in the dependent variable (here bacterial abundance, maximum − minimum = range) to this class width, that is

Received March 16, 1995. Accepted September 8, 1995.
J12820

Y.T. Prairie. Département des sciences biologiques, Université du Québec à Montréal, C.P. 8888, Station centre-ville, Montréal, QC H3C 3P8, Canada. e-mail: prairie.yves@uqam.ca

[1] A contribution to the Groupe de recherche interuniversitaire en limnologie (GRIL).

Fig. 1. Geometrical representation of the resolution power criterion. The number of risers on the staircase bounded by the 95% confidence intervals corresponds to the number of classes into which the dependent variable can be divided on the basis of prediction from the independent variable.



$$(1) \quad \text{Number}_{\text{classes}} = \frac{\text{RANGE}}{2(t_{(\text{df},p<0.05)}\sqrt{\hat{\sigma}_{\varepsilon}^{2}})}$$

The resolution power (i.e., the number of classes) is, as such, a simple and useful measure of the predictive utility of a model and eq. 1 can be applied directly to any regression analysis. For example, Bird and Kalff's (1984) model of bacterial abundance based on chlorophyll concentration (their eq. 3, r^2 = 0.88) has a resolution power of about 3.5 classes. Geometrically, the criterion of resolution power can be illustrated by simply constructing a staircase within the 95% confidence intervals (individual predictions) of the regression and counting the number of risers produced (Fig. 1).

The resolution power index is even more informative when compared with the familiar r^2 statistic. If one assumes that the observations are normally distributed along the axis of the ordinate and if one considers further that the whole range is effectively covered by the 99% confidence interval of the observations of the dependent variable, the resolution power can be calculated as

$$(2) \quad \text{Number}_{\text{classes}} = \frac{2t_{(\text{df},p<0.01)}\sqrt{\hat{\sigma}_{Y}^{2}}}{2t_{(\text{df},p<0.05)}\sqrt{\hat{\sigma}_{\varepsilon}^{2}}}$$

and since for any bivariate regression model

$$(3) \quad r^2 = 1 - \frac{(N-2)\,\hat{\sigma}_{\varepsilon}^{2}}{(N-1)\,\hat{\sigma}_{Y}^{2}}$$

substitution of eq. 3 in eq. 2 yields

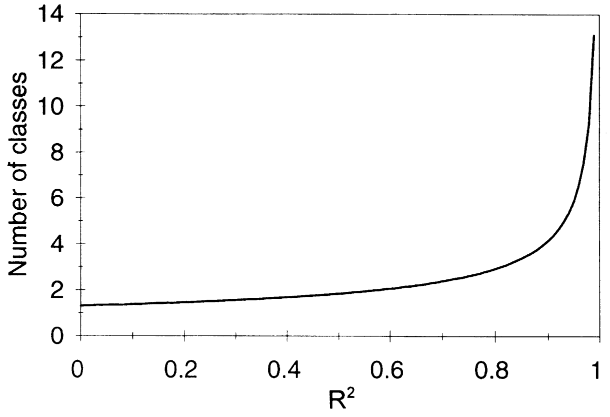
$$(4) \quad \text{Number}_{\text{classes}} = \frac{t_{(\text{df},p<0.01)}}{t_{(\text{df},p<0.05)}}\sqrt{\frac{(N-2)}{(N-1)(1-r^2)}}$$

which, for large N, reduces to

$$(5) \quad \text{Number}_{\text{classes}} \approx \frac{t_{(\infty,p<0.01)}}{t_{(\infty,p<0.05)}\sqrt{1-r^2}} \approx \frac{1.31}{\sqrt{1-r^2}}$$

Figure 2 illustrates the relationships, expressed by eq. 5, between the number of classes and the coefficient of determination (r^2). The striking feature of Fig. 2 is that the resolution power of a regression model increases very rapidly for r^2 > 0.65 but it remains surprisingly constant and low elsewhere. It requires an r^2 value of about 0.6 just to be able to statistically distinguish between only two classes. The rapidly increasing portion of the function for r^2 > 0.65 implies, however, that the search for additional explanatory variables is justified even when a single factor can account for a large fraction, say 90%, of the variance. Figure 2 shows that increasing the r^2 from 0.2 to 0.5 increases the resolution power by about 28%, whereas increasing the r^2 from 0.6 to 0.9 doubles it.

Fig. 2. The relationship between the number of distinguishable classes and the coefficient of determination (r^2).



Equation 5 expresses the general relationship between the number of classes and the r^2 but the exact value of the numerator depends on the stated assumptions regarding the distribution of the Y variable. For example, if we assume the observations of Y to be uniformly distributed (as opposed to normally distributed in the case above), the numerator takes the value of 0.88, which is obtained after substituting in eq. 3, the relationship between the range and variance of a rectangular variate ($\hat{\sigma}^2$ = range^2/12) and developing as in eqs. 3 to 5. Similarly, the calculations above use the constant EMS (σ_ε^2, eq. 1) as a measure of the uncertainty associated with an individual prediction when it is well known that prediction confidence limits widen away from the means (e.g., Kleinbaum et al. 1988). This widening would cause eq. 4 to overestimate the number of classes and hence the resolution power of the model. For N larger than 20, this effect can be shown to be negligible (Δnumerator < 5%). For small N, the exact solution becomes complicated but can be efficiently approximated (within 2–3%) by dividing the right-hand side of eq. 1 or 4 by the following correction factor (CF):

$$(6) \quad \text{CF} \approx \sqrt{(1+\frac{1}{N}) + \frac{(X_{\max}-\bar{X})^3 - (X_{\min}-\bar{X})^3}{3\hat{\sigma}_X^2\,(N-1)(X_{\max}-X_{\min})}}$$

where σ_X^2 is the estimated variance in X. In any case, it is the shape of the relationship between the r^2 value and the number of classes that is more significant here (Fig. 2), not the exact number of classes produced.

When and how should this measure of predictive power be used? The utility of this index is essentially conceptual in that it defines predictive power in intuitively simple terms: the number of separate classes into which the dependent variable can be divided. However, it is clear that it should not be used to estimate the boundaries of these categories, nor should the categories themselves be used in a predictive sense. Numerical predictions from standard regression statistics will always be more precise. Nevertheless, because this index can be translated in the mathematically rigorous statistic of r^2, it provides us with a tangible sense of what an r^2 value of say 0.9 actually means in simple terms. Sadly, however, given the ubiquity of published regression analyses with $r^2 < 0.65$, this exercise also shows that the bulk of aquatic sciences is still far from the desired power of resolution.

Acknowledgements

The author thanks Andrew Pawlisz for wanting to use this criterion and therefore forcing him to write it down. David Bird and two anonymous reviewers made valuable comments on the manuscript. This work was supported by an operating grant from the Natural Sciences and Engineering Research Council of Canada.

References

Bird, D.F., and Kalff, J. 1984. Empirical relationships between bacterial abundance and chlorophyll concentration in fresh and marine waters. Can. J. Fish. Aquat. Sci. **41**: 1015–1023.

Kleinbaum, D.G., Kupper, L.L., and Muller, K.E. 1988. Applied regression analysis and other multivariate methods. Duxbury Press, Belmont, Calif.