Model Evaluation by Comparison of Model-Based Predictions and Measured Values

Hugh G. Gauch, Jr., J. T. Gene Hwang, and Gary W. Fick*

ABSTRACT

The appropriateness of a statistical analysis for evaluating a model depends on the model's purpose. A common purpose for models in agricultural research and environmental management is accurate prediction. In this context, correlation and linear regression are frequently used to test or compare models, including tests of intercept \( a = 0 \) and slope \( b = 1 \), but unfortunately such results are related only obliquely to the specific matter of predictive success. The mean squared deviation (MSD) between model predictions \( X \) and measured values \( Y \) has been proposed as a directly relevant measure of predictive success, with MSD partitioned into three components to gain further insight into model performance. This paper proposes a different and better partitioning of MSD: squared bias (SB), nonunity slope (NU), and lack of correlation (LC). These MSD components are distinct and additive, they have straightforward geometric and analysis of variance (ANOVA) interpretations, and they relate transparently to regression parameters. Our MSD components are illustrated using several models for wheat (Triticum aestivum L.) yield. The MSD statistic and its components nicely complement correlation and linear regression in evaluating the predictive accuracy of models.

Comparisons of model-based and measured values arise frequently in agricultural research. For instance, a simulation model predicting yield from weather, soil, physiological, and morphological data can be compared with actual yield measurements to assess the model's accuracy and merit. Similarly, after a model has been developed using one data set, model outputs may be compared with a different data set used for validation. Typical purposes for such comparisons are to assess a model's predictive accuracy, to inform preferences among several competing models, to inform choices among various possible measurements serving as model inputs, to define the range of conditions over which a model is applicable or reliable, and to characterize the specific kinds of departures between model-based and measured values as a prelude to identifying specific possibilities for model refinement.

Here the convention is to use model outputs as predictors for actual measurements, symbolized by \( X \) and \( Y \), respectively. One simple statistic for assessing a model's merit is the correlation coefficient \( r \) between \( X \) and \( Y \). Another common analysis is linear regression of \( Y \) on \( X \) to check whether the intercept \( a \) is near 0 and the slope \( b \) is near 1. Wallach and Goffinet (1989) observed that agricultural and ecological researchers use diverse statistical analyses for model evaluation, but many of these analyses fail to quantify directly the predictive accuracy of a model, even when such is the researchers' explicit objective. This confusion persists. For instance, see the 10 papers from a symposium on "Crop Modeling and Genomics" published recently in this journal (Agronomy Journal 95:4–113). That symposium illustrates the frequent use of correlation and regression for model evaluation.

However, Kobayashi and Salam (2000) present cogent reasons why the correlation coefficient and linear regression are not entirely satisfactory for model evaluation and suggest that MSD and its components are often more informative. Further developing those findings, a different partitioning of MSD components has the advantage of yielding distinct components with straightforward meanings.

COMPONENTS OF MEAN SQUARED DEVIATION

Model-based and measured values, \( X \) and \( Y \), can be compared for the purpose of evaluating a simulation model. An important statistic for this purpose is the mean square deviation between \( X \) and \( Y \). The MSD is simply the sum of squared deviations between \( X \) and \( Y \), divided by the number of observations, \( N \), where summation is over \( n = 1, 2, \ldots, N \),

\[
MSD = \frac{\sum(X_n - Y_n)^2}{N} \quad [1]
\]

Let \( \bar{X} \) and \( \bar{Y} \) be the means. Also, let \( x_n = X_n - \bar{X} \) and \( y_n = Y_n - \bar{Y} \) be the deviations from the means.

The partitioning of MSD suggested by Kobayashi and Salam (2000) has three components [also see Xie et al. (2001) and Ewert et al. (2002)]. Using their notations, the first component is SB, which arises from these two means being unequal,

\[
SB = (\bar{X} - \bar{Y})^2 \quad [2]
\]

The (population) standard deviation of the simulation SD\( S \) is \( (\Sigma x_n^2 / N)^{1/2} \), and likewise the measurements have an SD\( M \) of \( (\Sigma y_n^2 / N)^{1/2} \). Accordingly, their second component, the difference in the magnitudes of fluctuation between the measurements and simulations (SDSD), arises from these two standard deviations being unequal,

\[
SDSD = (SD_M - SD_S)^2 \quad [3]
\]

Third and finally, there is lack of positive correlation weighted by the standard deviations (LCS).

Abbreviations: ANOVA, analysis of variance; LC, lack of correlation; LCS, lack of positive correlation weighted by the standard deviations of the measurements and simulations; MSD, mean squared deviation; MSEP, mean squared error of prediction; NU, nonunity slope; SB, squared bias; SDSD, difference in the magnitudes of fluctuation between the measurements and simulations.
LCS = 2SD,SD_{m}(1 - r) \quad [4]

However, Kobayasi and Salam (2000) note two problems with their MSD components. First, SDSD and LCS are not entirely independent since both involve SD, and SD_{m}. Consequently, their meanings are confounded and unclear. Second, it is difficult to use regression parameters (a,b, and r) in combination with MSD parameters (SB, SDSD, and LCS) for understanding data sets because these parameters “are not explicitly related to each other.” Also, a third problem could be mentioned, namely that their MSD components have no interpretation in terms of ANOVA (except, of course, for SB). Accordingly, a better partitioning of MSD was sought.

We retain SB since it is standard fare in statistics and has an entirely clear meaning. Incidentally, as a special case that b = 1, then SB > 0 if and only if the intercept a ≠ 0.

Our second component is the mean square for NU,

\[ NU = (1 - b)^2 \times \left( \Sigma x_i^2/N \right) \quad [5] \]

where \( b \) is the slope of the least-squares regression of Y on X, namely \( \Sigma x_i y_i/\Sigma x_i^2 \). Obviously, NU > 0 if and only if \( b \neq 1 \).

And our third component is the mean square for LC,

\[ LC = (1 - r)^2 \times \left( \Sigma y_i^2/N \right) \quad [6] \]

where \( r^2 \) is the square of the correlation, namely \( (\Sigma x_i y_i)^2/\Sigma x_i^2 \Sigma y_i^2 \). Obviously, LC > 0 if and only if \( r^2 \neq 1 \).

INTERPRETATION OF COMPONENTS

Our three components—SB, NU, and LC—add up to MSD; and they have distinct and clear meanings, including transparent relationships with regression parameters—NU with b and LC with \( r^2 \) (and even SB with a in the special case that b = 1). These components have a simple geometric interpretation that reinforces the meanings of their names, as shown in Fig. 1. The first panel shows the relationship of perfect equality, \( Y = X \), for which MSD = 0. Departure from perfect equality is possible in exactly three ways: as considered next: translation, rotation, and scatter.

The second panel of Fig.1 shows translation, the problem of SB resulting from \( X \neq Y \) (which in this special case with b = 1 emerges from a ≠ 0), for which SB = 1. The third panel (Fig. 1) shows rotation, NU resulting from \( b \neq 1 \), for which NU = 2. And finally, the fourth panel (Fig. 1) shows scatter, LC resulting from \( r^2 \neq 1 \) because of errors \( \varepsilon = (1, -1, 0) \), for which LC = 0.6667.

The resulting MSD for any combination of these three problems is simply additive. The model values are always the same, \( X = (-1, -1, 2) \). Combining NU and LC gives the data \( Y = 2X + \varepsilon = (-1, -3, 4) \), resulting in MSD = 2.6667 with components NU = 2 and LC = 0.6667. Likewise, combining all three problems gives \( Y = 1 + 2X + \varepsilon = (0, -2, 5) \), resulting in MSD = 3.6667 with the expected three components.

Incidentally, MSD itself also has a geometric interpretation. It is the mean square of the deviations around the 1:1 line in a plot of model predictions X against measured values Y, these deviations being reckoned as usual in the direction of the ordinate Y (although reckoning instead in the direction of the abscissa X necessarily gives identical deviations and the same result for MSD). And the square root of MSD is the (population) standard deviation of these deviations around the 1:1 line. For comparison, the root mean square error reported so commonly for regression analyses is the analogous (sample) standard deviation around the regression line (rather than the 1:1 line), having deviations reckoned in the direction of the ordinate for the usual regression of Y on X.

Because they are distinct and additive, our MSD components have simple and clear meanings. For comparison, the SB component is identical in the Kobayashi and Salam (2000) partitioning of MSD and in our partitioning, their SDSD component is roughly similar to our NU, and their LCS is roughly similar to our LC. But their SDSD and LCS are difficult to interpret. For example, for the fourth panel in Fig. 1 with LC = 0.6667, the Kobayashi and Salam (2000) analysis splits this simple LC into two components, SDSD = 0.0479 and LCS = 0.6188, so neither of these values has any obvious meaning. Likewise, when NU = 2 and LC = 0.6667 are combined as described in a previous paragraph, another awkward split occurs with SDSD = 2.3400 and LCS = 0.3267, which again has no transparent interpretation.

Another difference between the Kobayashi and Salam (2000) components of MSD and ours is that their components are unchanged by reversing X and Y, whereas ours do change. For example, consider reversing the axes.
in the last panel of Fig. 1, resulting in model \( X = (0, -2, 2) \) and data \( Y = (-1, -1, 2) \). Unlike the original regression with its slope of 1, this regression has a slope of \( b = 0.75 \), and since \( b \neq 1 \), consequently NU > 0, namely 0.1667. Furthermore, this new regression lies closer to its data points, resulting in the smaller LC = 0.5000. But given the symmetry in \( X \) and \( Y \) in Eq. [2], obviously SB is unchanged, namely 0. By contrast, all three of the Kobayashi and Salam (2000) components are symmetric and consequently unchanged. Again, reversal of \( X \) and \( Y \) has changed the slope and has changed the deviations from the regression line, and yet SDSD and LCS fail to reflect these important changes. Whether \( X \) or \( Y \) is the model, and correspondingly whether \( Y \) or \( X \) is the data, ought to matter for MSD components (except SB).

Besides their regression and geometric interpretations, our components also have an ANOVA interpretation. Consider the linear regression

\[
Y_n - X_n = a + bX_n + \varepsilon_n
\]

where \( \varepsilon_n \) is the error. The uncorrected total sum of squares, divided by \( N \), is MSD. The type I or sequential sum of squares for the intercept divided by \( N \) is SB. The type I sum of squares for \( X \), divided by \( N \), is NU. And finally, the sum of squares for error, divided by \( N \), is LC. Because of these facts, our partitioning of MSD has another nice property. Under the assumptions that \((X_n, \varepsilon_n)\) are independently identically distributed and that \(X_n\) and \(\varepsilon_n\) are independent, it can be shown that SB, NU, and LC are statistically independent (whereas SDSD and LCS are not). These independencies can be used to construct various \( F \) tests and \( t \) tests.

**EXAMPLE OF WHEAT MODELS**

Figure 2 shows SB, NU, and LC components of MSD for five wheat models, each compared with actual yield data from seven environments. The data are reported in Jamieson et al. (1998), and a comparable figure appears as Fig. 4 in Kobayashi and Salam (2000). Of course, MSD values are identical in these figures, as are SB values, but NU and SDSD differ somewhat, as do LC and LCS. For some models, the two analyses look rather similar, but for others there are greater differences. For instance, for the CERES model, NU is 0.1539, whereas the related SDSD of Kobayashi and Salam (2000) is only 0.0564, which differs by a factor of almost three. Another sizable discrepancy is that for SWHEAT, for which LC is 0.4174, but the related LCS is only 0.1590. Bear in mind, however, that regardless whether NU is numerically close to SDSD, those two statistics have different meanings, and the same applies to LC and LCS.

The striking feature of Fig. 2 is that the models differ greatly regarding which components contribute the most to MSD. The main problem for SUCROS2 is SB (41% of MSD), while for SWHEAT is NU (78%) and for CERES is LC (79%). Sirius has about equal proportions of all three problems, and AFRCWHEAT2 has mainly LC, some SB, and negligible NU.

The winner among these five wheat yield models according to MSD is AFRCWHEAT2. Coincidentally, this model also has the smallest LC component of MSD (as well as the smallest NU, although not the smallest SB). This model also achieves the highest correlation (0.9707). And for the usual regression of \( Y \) on \( X \), this model has the intercept closest to zero (0.3293), as well as the slope closest to one (1.0079). Hence, several considerations agree on AFRCWHEAT2 as being the best of these models for the seven environments used in this study.

Mean squared deviation components are best regarded as complements to regression parameters, rather than as replacements for them. Most pointedly, NU > 0 indicates that \( b \neq 1 \), but NU does not distinguish between \( b > 1 \) and \( b < 1 \), so knowing the slope provides important additional insight. This same verdict applies to the statistic that NU replaces, namely SDSD.

**DISCUSSION**

When the main purpose of a model is prediction, this imposes special interest in the 1:1 line of equality, \( Y = X \). By contrast, many applications of linear regression have no such special interest, such as a regression between yield and fertilizer level, for which there is no expectation whatsoever that the intercept be zero or that the slope be unity. Since the parameters in a linear re-
Regression account for bias and NU, only residual errors from LC reduce a regression’s fit. Accordingly, from the perspective of modeling for accurate predictions, this means that regression is responsive to only one of the three kinds of departures that degrade predictions: LC, but not SB and NU. In other words, MSD equates to deviations around the 1:1 line of equality, whereas the mean square error often reported for regressions uses deviations around the regression line.

The burden of Kobayashi and Salam (2000) is that MSD is better suited for model selection than is regression and correlation. A necessary precondition for this claim to be true is that MSD and $r$ rank models differently, at least sometimes—otherwise there is no difference in practice. For the present example of five wheat yield models, however, recall that AFRCWHEAT2 has both the smallest MSD and the highest $r$, so MSD and regression approaches select the same winner. So, when would MSD and $r$ give different rankings?

From Eq. [6], LC and $r^2$ are related inversely, so model rankings by (low) LC and rankings by (high) $r^2$ are identical. Accordingly, different winners will be selected if and only if one model has the lowest MSD but a different model has the lowest LC component of MSD (or equivalently, the highest $r$), which can happen if the lowest LC happens to be accompanied by relatively high SB or NU or both. Every difference and any potential superiority that MSD has over regression for purposes of model evaluation are due precisely to MSD considering not only LC but also SB and NU.

A special but important case arises when a model’s problems with SB and NU are relatively easy to fix or reduce, but not problems with LC. Then MSD ranks the models well as regards their current merits, but LC (or its correlate $r$) gives a better ranking of the potential merits of the models after relatively easily fixed defects have been corrected. Unfortunately, failure to distinguish between current and potential merits could result in hasty dismissal of a promising model.

In the present context of the data $Y$ having the special role of being the standard of comparison, MSD is related to the mean squared error of prediction (MSEP), which is a standard statistic for assessing predictive accuracy (Wallach and Goffinet, 1989). By definition, MSEP is the mean squared difference between model predictions and the true values, which can be estimated from available empirical data by means of proper calculations even though the true values are inaccessible theoretical quantities. Hence, MSEP reflects but one kind of imperfection, the model’s imperfection, which clearly is the single most informative statistic regarding a model’s predictive accuracy. By contrast, MSD has two sources of discrepancies between model $X$ and data $Y$, namely that $X$ is not perfect, and neither is $Y$.

In the special but moderately frequent case that the errors in $Y$ are small relative to those in $X$, MSD is a close surrogate for MSEP. Otherwise, MSD needs to be discounted for the discrepancies originating from the validation data $Y$ to estimate the errors in the model $X$ itself, MSEP (Gauch, 1992, p. 134–153, especially page 140; Gauch, 2003, p. 303–312). In either case, the relationship between MSD and MSEP explains in part the inherent importance of MSD, especially when the data $Y$ have relatively small errors. Even if the magnitude of the errors in the validation data cannot be quantified accurately (because of lack of replication) so that MSEP values cannot be calculated, and even though corresponding MSEP and MSD values will differ, nevertheless the rankings of models by MSD can be expected to be similar to those that would emerge from MSEP. In such situations, MSD values (or more precisely, MSD rankings) are still useful for selecting the most predictively accurate model.

There are interesting relationships between regression and MSD components (beyond the transparent connections already noted between NU and $b$ and between LC and $r$). Consider the least-squares regression estimator for the dependent variable $Y_n$, which is ordinarily denoted by $\hat{Y}_n$, but here more conveniently denoted by $Z_n$, namely $Z_n = a + bX_n$. The vector, $Z$, opens up two new possibilities for MSD calculations: $Z$ could replace $Y$ or else $X$.

First, for the MSD comparison of $X$ and $Z$, replacement of $Y$ by $Z$ is equivalent geometrically to projecting points vertically onto the regression line. Accordingly, all scatter has been eliminated, so $r^2 = 1$ and hence LC = 0. But SB and NU remain the same as they were for the original comparison of $X$ and $Y$.

Second, for the MSD comparison of $Z$ and $Y$, replacement of $X$ by $Z$ is equivalent geometrically to projecting points horizontally onto the 1:1 line of equality. The regression line is automatically the 1:1 line, so $a = 0$ and $b = 1$ and hence SB = 0 and NU = 0. But LC remains the same.

Note that these two replacements have clear and simple consequences for MSD components, which are exact opposites. The first replacement could be of special interest when the LC component of MSD is large, whereas the second could be of interest when the SB or NU component or both components are large.

The usual reason for interest in $Z$, related to this first replacement, is that regression estimates are often closer to the true values than are the imperfect data $Y$. Basically, accuracy and efficiency are gained because parsimonious models can reduce noise and because regressions are based on more data than just one individual datum (Gauch, 1993, 2003, p. 291–296; Gruber, 1998).

But this second replacement is also interesting. Applying the trivial transformation $a + bX_n$ to model outputs $X_n$ can be regarded as a patch to the model that automatically eliminates the SB and NU components of MSD. For instance, for the five wheat yield models shown in Fig. 2, on average this simple patch reduces MSD to only 46% of the original values.

The real issue meriting more research, however, is not how much this patch helps for model outputs and data used in constructing this patch, but rather how much this patch would help for other (or new) model outputs and data not used in constructing this patch. Needless to say, merely patching a model and properly refining a model are two different matters. But sometimes a quick patch may have some utility, especially if
a validation study has proven that it improves predictive accuracy. Furthermore, awareness of MSD components may assist and accelerate subsequent model refinement.

In conclusion, MSD and its components are useful for model evaluation, including assessing predictive accuracy (which implies special interest in the 1:1 line of equality). But MSD measures discrepancies between model predictions and validation observations, whereas the more direct measure of predictive accuracy is MSEP, which reflects discrepancies between model predictions and true values. Accordingly, MSD is more valuable when the errors in the validation data $Y$ are rather small compared with the errors in the model predictions $X$, in which case MSD approaches MSEP. Regression parameters are also informative, particularly since the slope $b$ is closely related to the NU component of MSD and the squared correlation $r^2$ is closely related to LC, as Eq. [5] and [6] show.

ACKNOWLEDGMENTS

This research was supported in part by Hatch funding allocated to the Cornell Agricultural Experiment Station and to Regional Research Project NE-132. We appreciate insightful suggestions from the anonymous reviewers, particularly the analysis-of-variance interpretation of our MSD components.

REFERENCES


