Article

Patterns

Bias-variance decomposition of absolute errors for diagnosing regression models of continuous data

Highlights

- The bias-variance decomposition (BVD) of absolute error was analytically derived
- Absolute and squared errors view bias/variance trade-off and ensembles differently
- Absolute error better illustrates relative importance of various BVD error sources
- Absolute-error BVD better promotes model traits that reduce estimation residuals

Authors

Jing Gao

Correspondence

jinggao@udel.edu

In brief

Regression models of continuous data are widely used machine learning techniques in science and engineering. Properly defined, appropriately chosen error metrics are necessary for understanding and improving these models and for using their results to support decision making. This work analytically derives the bias-variance decomposition (BVD) of absolute error and compares it with the commonly used squared-error BVD. The results encourage the profession to transition from habitually using squared errors to adopting absolute errors and highlight the strengths of BVD for studying data-driven models.

Gao, 2021, Patterns *2*, 100309
August 13, 2021
https://doi.org/10.1016/j.patter.2021.100309

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Bias-variance decomposition of absolute errors for diagnosing regression models of continuous data

Jing Gao[1,2,*]
[1]Data Science Institute & Department of Geography and Spatial Sciences, University of Delaware, Newark, DE 19716, USA
[2]Lead contact
*Correspondence: jinggao@udel.edu
https://doi.org/10.1016/j.patter.2021.100309

THE BIGGER PICTURE As machine learning becomes the foundation of many scientific and engineering models, model development is increasingly automated, where once model success can be clearly defined (e.g., using estimation error metrics), the machine can learn underlying patterns and construct successful models on its own through computational trial and error. Properly defined, appropriately chosen error metrics therefore hold a key to unlock the power of data-driven modeling. This work introduces new error metrics attributing estimation errors to different sources and highlights unique strengths of the new technique relative to existing methods. The results inform effective use of error metrics for understanding and improving models of continuous data. The work is timely in the context of rising debates about uncertainty, reproducibility, and transparency in data science. Ultimately, better error metrics can lead to better models and better artificial intelligence systems for greater societal benefits.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Bias-variance decomposition (BVD) is a powerful tool for understanding and improving data-driven models. It reveals sources of estimation errors. Existing literature has defined BVD for squared error but not absolute error, while absolute error is the more natural error metric and has shown advantages over squared error in many scientific fields. Here, I analytically derive the absolute-error BVD, empirically investigate its behaviors, and compare that with other error metrics. Different error metrics offer distinctly different perspectives. I find the commonly believed bias/variance trade-off under squared error is often absent under absolute error, and ensembles—a never hurt technique under squared error—could harm performance under absolute error. Compared with squared error, absolute-error BVD better promotes model traits reducing estimation residuals and better illustrates relative importance of different error sources. As data scientists pay increasing attention to uncertainty issues, the technique introduced here can be a useful addition to a data-driven modeler's toolset.

INTRODUCTION

Model evaluation and error assessment provide necessary information for using data-driven modeling results as a basis for real-world decision making. Most existing model evaluation methods and metrics focus on describing model performance (e.g., the commonly used root mean square error [RMSE] describes how much estimation error is present), while more useful for the analyst are methods that could inform model improvement efforts, e.g., revealing what model characteristics contribute how much estimation error. The bias-variance decomposition (BVD) is a tool of this kind for diagnosing data-driven models.

The BVD attributes the expected estimation error of a modeling method at every data point to three model characteristics: bias, variance, and noise (precise mathematical definitions are provided in the analytical results). These three model characteristics are independent from each other, and so are their contributions to the expected estimation error, which are defined as

bias effect, variance effect, and noise effect. The distinctions between "characteristics" and "effects" are important[1]: for example, "variance" (a characteristic quantity) measures the sensitivity of a model to training data, and "variance effect" (an effect quantity) is the amount of estimation error caused by this sensitivity. Because the three model "characteristics" are independent from each other and respond to different model alteration techniques, the expected estimation error is the sum of the three "effect" quantities, and understanding how much each source contributes to expected estimation error can effectively inform model improvement efforts.

The relationship between the three model characteristics and their effects is a function of how error is defined. BVD was initially developed for squared error (i.e., error is the squared value of the difference between observation and estimation), where BVD model characteristics and their effects equate, and expected error is simply the sum of bias, variance, and noise.[2] Due to its mathematical tractability, squared error is the most popular error metric in today's model evaluation and error assessment practices. Meanwhile, squared-error BVD's conceptual simplicity and explanatory power inspired the development of a generalized BVD definition, which was then applied to the popular classification error metric—zero-one error (i.e., error is one if misclassified and zero if correctly classified).[3] Because of the intrinsic mathematical differences between zero-one and squared errors, the BVD model characteristics and effects exhibit more complex relations under zero-one error. Nonetheless, by revealing how BVD model characteristics cast their effects on estimation error, zero-one-error BVD has proven a powerful tool for diagnosing data-driven classification models.

In many scientific fields that utilize data-driven models of continuous response variables (i.e., regressions, broadly defined), absolute error (i.e., error is the absolute value of the difference between observation and estimation) is the most natural error metric. For example, it is more intuitive to interpret a population growth model's performance, if error is measured as the magnitude of over- or under-estimation in the unit of people rather than people-squared. For another example, climate modelers have reported advantages of using absolute error over squared error in assessing climate model performance.[4,5] Data-driven modeling efforts in these fields can benefit from absolute-error BVD analyses.

However, few attempts have been made at defining BVD for absolute error. On the one hand, analytically deriving absolute-error BVD is a challenging task, due to absolute error's lack of many convenient mathematical properties of squared error. On the other hand, although methodologists understand the significance of contrasting absolute versus squared errors, domain scientists might be inclined to use readily available tools from existing software, which may reinforce the prevalent convention of using squared error in places where absolute error is more appropriate. James[1] proposed an absolute-error BVD definition. Although he conceptually emphasized the importance of distinguishing "characteristics" and "effects," the proposed definition treats the two sets of quantities independently, which contradicts the concept that effects arise from characteristics. Under this definition,[1] some empirical correlation is often found between corresponding characteristic and effect, but the nature of the relation is arbitrary. The definition also adds false dependency among bias, variance, and noise effects, and does not guarantee unique solutions.

Here, I derive the analytical solution of absolute-error BVD, by applying the widely accepted generalized definitions of BVD characteristics (bias, variance, noise) to the mathematical definition of absolute error. The resulting absolute-error BVD definition is unique, and naturally maintains the independence among bias, variance, and noise effects. The relationships between the three BVD model characteristics and their respective effects naturally arise from the analytical deduction. I then apply this absolute-error BVD to an example data point, to illustrate, at the individual data point level, how absolute-error and squared-error BVD compare. I also systematically test the empirical behaviors of the absolute-error BVD for evaluating regression modeling methods, using 11 benchmark datasets from the University of California Irvine (UCI) Machine Learning Repository. I examined a number of regression modeling methods: k-nearest neighbor induction, support vector machine (SVM) regression with linear kernel (SVM-linear), SVM regression with Gaussian kernel (SVM-Gauss), regression tree, and bagging using regression trees (for understanding absolute-error BVD's reaction to ensembles). I used Weka 3's standard implementation for all algorithms and varied the model complexity control parameter for each regression modeling method to investigate how absolute-error BVD components react to changes in model complexity. Below, in the analytical results, I present the statistical underpinning of the analytically derived absolute-error BVD along with the example data point analysis, and in the empirical results key findings of the empirical experiments, while detailed descriptions of the experiments and model-by-model results are in the supplemental experimental procedures. Considering that squared error is likely the most familiar error metric to many readers, I contrast the absolute-error results with corresponding squared-error BVD analyses as a point of reference for interpretations.

This absolute-error BVD definition has been applied to diagnose and improve a real-world geospatial model of satellite image processing for mapping urban land development.[6] The in-depth case study highlighted the method's usefulness generalizable to other data-driven modeling applications. The BVD acknowledges that model estimation errors at different data points could follow different probability distributions (Figure 1B), in contrast to the typical aggregate error assessment that assumes model estimation errors at all data points follow the same probability distribution and only analyzes aggregate error metrics (Figure 1A). When conducted for every data point, BVD can reveal the relative importance of different error sources in different parts of the study area, and the analyst can effectively improve model performance by applying different methods (targeting bias or variance) to different parts of the study area.

Altogether, data-driven modelers who are interested in understanding where estimation errors come from, how the error composition varies across the feature space, and how to efficiently design model improvement strategies, would likely find this absolute-error BVD a useful addition to their toolset.

Analytical results

Generalized BVD definition

Below, I introduce our notations through a brief summary of the widely accepted generalized BVD definition[1–3].

In modeling, analysts aim to estimate the response variable Y using the predictor (vector) $\boldsymbol{X}$, with a model (also known as

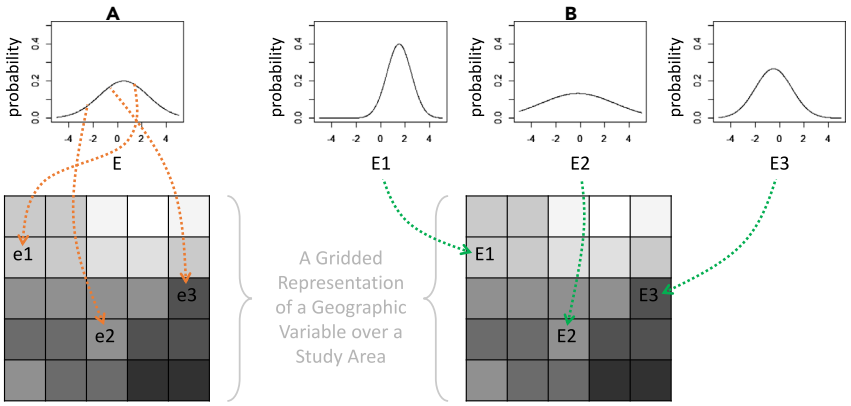

Figure 1. Typical aggregate error assessment versus BVD for individual data points

(A) Typical aggregate error assessment, assuming model estimation errors at all data points follow the same probability distribution, versus, (B) BVD for individual data points, acknowledging model estimation errors at different data points could follow different probability distributions. Upper-cased symbols (e.g., E) denote random variables, and lowercase symbols (e.g., e1) denote individual values of a random variable.

function, or machine) f constructed by a learning algorithm (also known as modeling method) using a collection of observed $(\boldsymbol{X}, Y)$ pairs, i.e., a training set $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x_n}, y_n)$, so that the model estimation $\hat{Y} = f(\boldsymbol{X})$ approximates the observed response Y. The estimated $\hat{Y}$ may be referred to as fitted values for training data points, and predicted values for test data points (i.e., data that are used to evaluate the model and are independent from training). While the BVD can be applied to both fitted and predicted values in the same way, below we focus on data points used for model evaluation and refer to values of $\hat{Y}$ as predictions. Here, upper-cased symbols denote random variables, and lower-cased symbols individual values of a random variable; bold symbols denote vectors of variables (if uppercase) or values (if lowercase), and non-bold symbols denote individual variables or values.

An error definition (also known as loss function) $L(Y, \hat{Y})$ measures the cost of predicting $\hat{Y}$ when the observation is Y. Commonly used loss functions include squared error $L_{sq}(Y, \hat{Y}) = (Y - \hat{Y})^2$ and absolute error $L_{abs}(Y, \hat{Y}) = \left|Y - \hat{Y}\right|$ for regressions, and zero-one error for classifications: $L_{01}\left(Y, \hat{Y}\right) = 0$ *if* $Y = \hat{Y}$, *and* $L_{01}\left(Y, \hat{Y}\right) = 1$ *otherwise*.

The generalized BVD framework applies to any loss function whose expectation is computable. It attributes the expected error of a modeling method at a data point $\boldsymbol{x}$ with respect to a set D of varying training sets and the probability distribution of the observed response Y conditional on $\boldsymbol{x}$, i.e., $E_{D,Y}[L(Y, \hat{Y})]$, to three independent sources—bias, variance, and noise. Since BVD addresses individual data points, all terms defined below are conditioned on the event $\{\boldsymbol{X} = \boldsymbol{x}\}$. To simplify, this conditional clause will not be written out, but readers should assume it is always in effect unless otherwise noted.

The generalized BVD is defined using optimal prediction and main prediction. The optimal prediction y_o of a data point $\boldsymbol{x}$ is the central tendency of the probability distribution of all the observed values of Y for that $\boldsymbol{x}$, i.e., $y_o = argmin_{y_o} E_Y[L(Y, y_o)]$. It is the best possible guess of Y given $\boldsymbol{x}$. The main prediction $\hat{y}_m$ of a data point $\boldsymbol{x}$ is the central tendency of the probability distribution of all the predictions generated for $\boldsymbol{x}$ by various models constructed using different training sets in D, i.e., $\hat{y}_m = argmin_{\hat{y}_m} E_D[L(\hat{Y}, \hat{y}_m)]$. It is the prediction that minimizes the expected deviation of $\hat{Y}$ from $\hat{y}_m$ with respect to the probability distribution of predictions made for $\boldsymbol{x}$. y_o and $\hat{y}_m$ are both central tendency measures (Figure 3), while the meaning of "central" depends on what loss function is used, e.g., the main prediction of $\boldsymbol{x}$ under squared error is the mean of all predictions for $\boldsymbol{x}$; under absolute error, it is the median; and under zero-one error, the mode.

The bias of a modeling method at a data point $\boldsymbol{x}$ is the difference between $\boldsymbol{x}$'s optimal prediction and main prediction, $B = L(y_o, \hat{y}_m)$. Bias captures the systematic difference between the optimal model and the model at hand (Figure 2). Bias effect, BE, is the amount of estimation error caused by this difference. Bias often arises when the model's assumptions do not match with the observed patterns in data. For example (Figure 3A), if the observations of two random variables show that Y is quadratic to X but the analyst chose to model the data with a linear regression, the misassumption of linearity will make the resulting model systematically under-estimate at the high and the low ends of the range of X and over-estimate at the mid-range of X. Such systematic effect would be attributed to bias.

The variance of a modeling method at a data point $\boldsymbol{x}$ is the expected deviation of a model prediction from the main prediction with respect to all training sets in D, $V = E_D[L(\hat{y}_m, \hat{Y})]$. Variance is independent of the observed response, and captures the sensitivity of a modeling method to small variations in training data (Figure 2). Variance effect, VE, is the amount of estimation error caused by this sensitivity. Variance arises when different models result from different yet all adequate training sets used to train the same modeling method. For example, in Figure 3B, 1,000 linear regression models (each shown as a gray line) were, respectively, trained using 1,000 independently sampled training sets about the same phenomenon. The variation across all the gray lines is the variance of the modeling method (i.e., linear regression here).

Noise is the variability of Y given $\boldsymbol{x}$, $N = E_Y[L(Y, y_o)]$ (Figure 2). Noise is independent of the modeling method, and arises from the observational setup, where different values of Y can be observed for the same $\boldsymbol{x}$ (Figure 3C). Noise effect, NE, captures the collective impact of all the factors that may influence the observed response but are not included in the model. Noise (effect) is unavoidable and usually unattainable, and hence is often assumed to be zero or simply ignored in practical applications.

Applying these definitions to squared error, one can deduce for a data point $\boldsymbol{x}$ that $E_{D,Y}[L_{sq}(Y, \hat{Y})] = BE + VE + NE = B + V + N$, where $B = (y_o - \hat{y}_m)^2$, $V = E_D[(\hat{y}_m - \hat{Y})^2]$, and $N = E_Y[(Y - y_o)^2]$. Under squared error, BVD "characteristics" and "effects" equate. Squared-error BVD has been well studied for

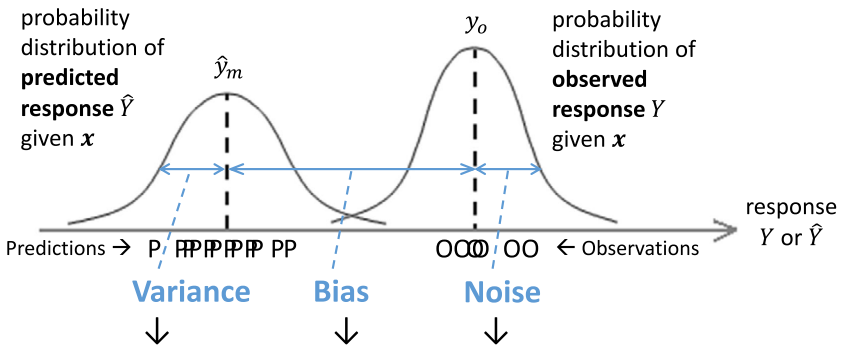

Figure 2. Generalized BVD definitions of bias, variance, noise, and their effects

This is a figurative graph. Normality is not required for BVD analyses.

explaining and improving model performance. To reduce squared-error variance, ensemble methods have proven effective and never hurt performance.[7] To reduce bias calls for the structure of the model to better fit the nature of the problem. There is a trade-off between squared-error bias and variance that prevents effective concurrent reduction of both error components (i.e., the bias/variance trade-off). To overcome this challenge, specialized models with innate assumptions that reflect accurate a priori knowledge about the problem being modeled are recommended over non-parametric models.[2]

Applying the generalized BVD definitions to zero-one error, one can deduce for a data point $\boldsymbol{x}$ that $E_{D,Y}[L_{01}(Y,\hat{Y})] = BE + VE + NE = B + cV + (P_D(\hat{Y} = y_o) - P_D(\hat{Y} \neq y_o)P_Y(\hat{Y} = Y \mid Y \neq y_o))N$, where $B = 0$ if $\hat{y}_m = y_o$, and $B = 1$ otherwise; $c = 1$ if $\hat{y}_m = y_o$, and $c = -P_D(\hat{Y} = y_o \mid \hat{Y} \neq \hat{y}_m)$ otherwise; $V = P_D(\hat{Y} \neq \hat{y}_m)$; $N = P_Y(Y \neq y_o)$. Under zero-one error, $BE = B$, but parts of V and N reduce rather than increase expected error, because certain types of variations can shift already biased classifications into correct categories. As a result, ensembles, which reduce all variations with no regard to their effects, can transform poorly performing classification models into worse ones.[7,8] Also, zero-one error bias and variance do not always show trade-off.[3] Altogether, zero-one error is considered more forgiving than squared error, as not all BVD characteristics increase expected zero-one error.[3,9]

Analytical derivation of absolute-error BVD

Because bias, variance, and noise are independent of each other, so are their effects. That is, $E_{D,Y}[L(Y,\hat{Y})] = BE + VE + NE$ always hold, but how the effects relate to the characteristics (i.e., B, V, and N) vary for different loss functions.

Proposition 1. Applying the generalized definition of BVD characteristics to absolute error, one will have $L_{abs}(Y,\hat{Y}) = |Y - \hat{Y}|$, bias $B = |y_o - \hat{y}_m|$, variance $V = E_D[|\hat{Y} - \hat{y}_m|]$, and noise $N = E_Y[|Y - y_o|]$. Then, expected absolute error at a given data point $\boldsymbol{x}$ is the sum of bias effect BE, variance effect VE, and noise effect NE, i.e., $E_{D,Y}[L_{abs}(Y,\hat{Y})] = BE + VE + NE$, with

$BE = (P_{D,Y}(s_{biasEff} = 1) - P_{D,Y}(s_{biasEff} = -1))B$, where the sign coefficient of bias effect $s_{biasEff} = -1$ *if* $(Y \geq \hat{Y}$ *and* $y_o < \hat{y}_m)$ *or* $(Y < \hat{Y}$ *and* $y_o > \hat{y}_m)$, and $s_{biasEff}=1$ otherwise;

$VE = V - 2E_{D,Y}[L_{abs}(\hat{Y},\hat{y}_m) \mid s_{varEff} = -1]P_{D,Y}(s_{varEff} = -1)$, where the sign coefficient of variance effect $s_{varEff} = -1$ *if* $(Y \geq \hat{Y}$ *and* $\hat{Y} > \hat{y}_m)$ *or* $(Y < \hat{Y}$ *and* $\hat{Y} < \hat{y}_m)$, and $s_{varEff}=1$ otherwise;

$NE = N - 2E_{D,Y}[L_{abs}(Y, y_o) \mid s_{noiseEff} = -1]P_{D,Y}(s_{noiseEff} = -1)$, where the sign coefficient of noise effect $s_{noiseEff} = -1$ *if* $(Y \geq \hat{Y}$ *and* $Y < y_o)$ *or* $(Y < \hat{Y}$ *and* $Y > y_o)$, and $s_{noiseEff}=1$ otherwise.

Proof. To decompose $E_{D,Y}[L_{abs}(Y,\hat{Y})]$, let us first look at the quantity $L_{abs}(y_o,\hat{Y}) = |y_o - \hat{Y}|$. Taking $y_o - \hat{Y}$ out of the absolute-value operator results in two possible cases:

$$L_{abs}(y_o,\hat{Y}) = \begin{cases} y_o - \hat{Y} = (y_o - \hat{y}_m) + (\hat{y}_m - \hat{Y}), & \hat{Y} \leq y_o \quad (1) \\ -(y_o - \hat{Y}) = -(y_o - \hat{y}_m) - (\hat{y}_m - \hat{Y}), & \hat{Y} > y_o \quad (2) \end{cases}$$

For case (1):

$$L_{abs}(y_o,\hat{Y}) = y_o - \hat{Y} = \begin{cases} |y_o - \hat{y}_m| + |\hat{y}_m - \hat{Y}| = L_{abs}(y_o,\hat{y}_m) + L_{abs}(\hat{y}_m,\hat{Y}), & y_o \geq \hat{y}_m \text{ and } \hat{Y} \leq \hat{y}_m \\ L_{abs}(y_o,\hat{y}_m) - L_{abs}(\hat{y}_m,\hat{Y}), & y_o \geq \hat{y}_m \text{ and } \hat{Y} > \hat{y}_m \\ -L_{abs}(y_o,\hat{y}_m) + L_{abs}(\hat{y}_m,\hat{Y}), & y_o < \hat{y}_m \text{ and } \hat{Y} \leq \hat{y}_m \end{cases}$$

The condition ($y_o < \hat{y}_m$ and $\hat{Y} > \hat{y}_m$) contradicts with the condition of case (1) $\hat{Y} \leq y_o$, so will never occur.

For case (2):

$$L_{abs}(y_o,\hat{Y}) = -(y_o - \hat{Y}) = \begin{cases} -|y_o - \hat{y}_m| + |\hat{y}_m - \hat{Y}| = -L_{abs}(y_o,\hat{y}_m) + L_{abs}(\hat{y}_m,\hat{Y}), & y_o > \hat{y}_m \text{ and } \hat{Y} \geq \hat{y}_m \\ L_{abs}(y_o,\hat{y}_m) - L_{abs}(\hat{y}_m,\hat{Y}), & y_o \leq \hat{y}_m \text{ and } \hat{Y} < \hat{y}_m \\ L_{abs}(y_o,\hat{y}_m) + L_{abs}(\hat{y}_m,\hat{Y}), & y_o \leq \hat{y}_m \text{ and } \hat{Y} \geq \hat{y}_m \end{cases}$$

The condition ($y_o > \hat{y}_m$ and $\hat{Y} < \hat{y}_m$) contradicts with the condition of case (2) $\hat{Y} > y_o$, so will never occur.

In summary, $L_{abs}(y_o,\hat{Y}) = c_1 L_{abs}(y_o,\hat{y}_m) + c_2 L_{abs}(\hat{y}_m,\hat{Y})$, *where* $c_1 = -1$ *if* $(\hat{Y} \leq y_o$ *and* $y_o < \hat{y}_m)$ *or* $(\hat{Y} > y_o$ *and* $y_o > \hat{y}_m)$,

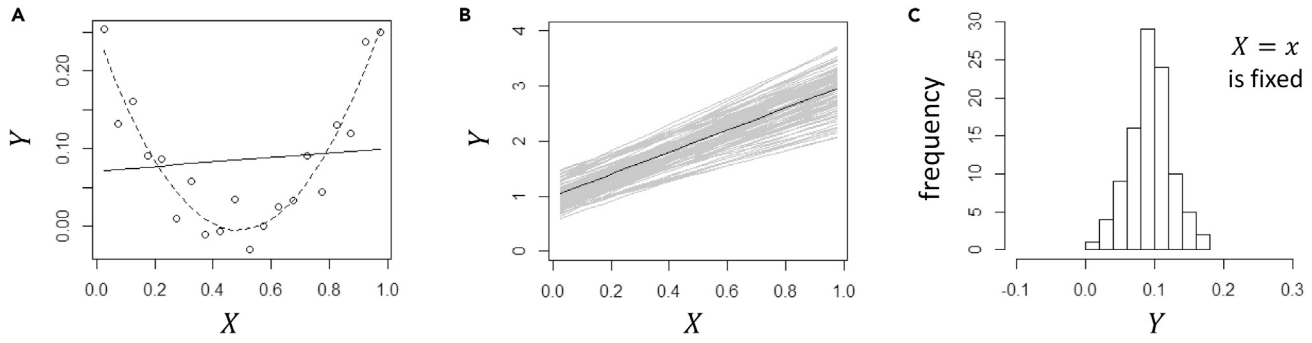

Figure 3. Example sources of bias, variance, and noise

Example sources of (A) bias, (B) variance, and (C) noise. The three figure panels (A), (B), and (C) are not related, and each is a figurative example of how the corresponding BVD error component might arise: (A) when the observed data points (small circles) show that Y is quadratic to X (dashed line) but the analyst fits the relation with a linear model (solid line), the systematic under-estimation at the high and the low ends of the range of X and the over-estimation at the mid-range of X contribute to bias (effect); (B) the black line shows the true relation between Y and X, while 1,000 linear regression models (each shown as a gray line) were, respectively, trained using 1,000 independently sampled training datasets, and the variation across the gray lines is the variance of the modeling method (i.e., linear regression here); (C) a histogram of the different Y values observed for the same x, and the variation of the distribution is the noise at the data point x.

and $c_1 = 1$ *otherwise*; $c_2 = -1$ *if* $\left(\widehat{Y} \le y_o \text{ and } \widehat{Y} > \widehat{y}_m\right)$ *or* $\left(\widehat{Y} > y_o \text{ and } \widehat{Y} < \widehat{y}_m\right)$, *and* $c_2 = 1$ *otherwise*.

Similarly, $L_{abs}(Y, \widehat{Y}) = c_{01}L_{abs}(Y, y_o) + c_{02}L_{abs}(y_o, \widehat{Y})$, *where* $c_{01} = -1$ *if* $(Y \ge \widehat{Y}$ *and* $Y < y_o)$ *or* $(Y < \widehat{Y}$ *and* $Y > y_o)$, *and* $c_{01} = 1$ *otherwise*; $c_{02} = -1$ *if* $(Y \ge \widehat{Y}$ *and* $\widehat{Y} > y_o)$ *or* $(Y < \widehat{Y}$ *and* $\widehat{Y} < y_o)$, *and* $c_{02} = 1$ *otherwise*.

Combining these two equations, we have $L_{abs}(Y, \widehat{Y}) = c_{01}L_{abs}(Y, y_o) + c_{02}c_1L_{abs}(y_o, \widehat{y}_m) + c_{02}c_2L_{abs}(\widehat{y}_m, \widehat{Y})$.

To simplify, $L_{abs}(Y, \widehat{Y}) = s_{noiseEff}L_{abs}(Y, y_o) + s_{biasEff}L_{abs}(y_o, \widehat{y}_m) + s_{varEff}L_{abs}(\widehat{y}_m, \widehat{Y})$, *where* $s_{noiseEff} = -1$ *if* $(Y \ge \widehat{Y}$ *and* $Y < y_o)$ *or* $(Y < \widehat{Y}$ *and* $Y > y_o)$, *and* $s_{noiseEff} = 1$ *otherwise*; $s_{biasEff} = -1$ *if* $(Y \ge \widehat{Y}$ *and* $y_o < \widehat{y}_m)$ *or* $(Y < \widehat{Y}$ *and* $y_o > \widehat{y}_m)$, *and* $s_{biasEff} = 1$ *otherwise*; $s_{varEff} = -1$ *if* $(Y \ge \widehat{Y}$ *and* $\widehat{Y} > \widehat{y}_m)$ *or* $(Y < \widehat{Y}$ *and* $\widehat{Y} < \widehat{y}_m)$, *and* $s_{varEff} = 1$ *otherwise*. Clearly, $s_{noiseEff}$, $s_{biasEff}$, and s_{varEff} are all functions of Y and $\widehat{Y}$.

Let us now analyze the expectation of $L_{abs}(Y, \widehat{Y})$ with respect to the probability distribution of observed Y given $\boldsymbol{x}$ and the training sets in D. We then have the following linear decomposition:

$$\begin{aligned} E_{D,Y}\left[L_{abs}\left(Y, \widehat{Y}\right)\right] &= E_{D,Y}\left[s_{noiseEff}L_{abs}(Y, y_o) + s_{biasEff}L_{abs}\left(y_o, \widehat{y}_m\right) + s_{varEff}L_{abs}\left(\widehat{y}_m, \widehat{Y}\right)\right] \\ &= E_{D,Y}[s_{noiseEff}L_{abs}(Y, y_o)] + E_{D,Y}[s_{biasEff}]L_{abs}\left(y_o, \widehat{y}_m\right) + E_{D,Y}\left[s_{varEff}L_{abs}\left(\widehat{y}_m, \widehat{Y}\right)\right] \end{aligned}$$

The first term of the above equation describes the overall effect of noise:

$$\begin{aligned} E_{D,Y}[s_{noiseEff}L_{abs}(Y, y_o)] &= E_{D,Y}[L_{abs}(Y, y_o)|s_{noiseEff} = 1]P_{D,Y}(s_{noiseEff} = 1) \\ &\quad - E_{D,Y}[L_{abs}(Y, y_o)|s_{noiseEff} = -1]P_{D,Y}(s_{noiseEff} = -1) \\ &= E_Y[L_{abs}(Y, y_o)] - 2E_{D,Y}[L_{abs}(Y, y_o)|s_{noiseEff} = -1]P_{D,Y}(s_{noiseEff} = -1) \\ &= N - 2E_{D,Y}[L_{abs}(Y, y_o)|s_{noiseEff} = -1]P_{D,Y}(s_{noiseEff} = -1) = NE \end{aligned}$$

The second term describes the overall effect of bias:

$$\begin{aligned} E_{D,Y}[s_{biasEff}]L_{abs}\left(y_o, \widehat{y}_m\right) &= (P_{D,Y}(s_{biasEff} = 1) - P_{D,Y}(s_{biasEff} = -1))L_{abs}\left(y_o, \widehat{y}_m\right) \\ &= (P_{D,Y}(s_{biasEff} = 1) - P_{D,Y}(s_{biasEff} = -1))B = BE \end{aligned}$$

The third term describes the overall effect of variance:

$$\begin{aligned} E_{D,Y}\left[s_{varEff}L_{abs}\left(\widehat{y}_m, \widehat{Y}\right)\right] &= E_{D,Y}\left[L_{abs}\left(\widehat{y}_m, \widehat{Y}\right)\middle| s_{varEff} = 1\right]P_{D,Y}(s_{varEff} = 1) \\ &\quad - E_{D,Y}\left[L_{abs}\left(\widehat{y}_m, \widehat{Y}\right)\middle| s_{varEff} = -1\right]P_{D,Y}(s_{varEff} = -1) \\ &= E_D\left[L_{abs}\left(\widehat{y}_m, \widehat{Y}\right)\right] - 2E_{D,Y}\left[L_{abs}\left(\widehat{y}_m, \widehat{Y}\right)\middle| s_{varEff} = -1\right]P_{D,Y}(s_{varEff} = -1) \\ &= V - 2E_{D,Y}\left[L_{abs}\left(\widehat{y}_m, \widehat{Y}\right)\middle| s_{varEff} = -1\right]P_{D,Y}(s_{varEff} = -1) = VE \end{aligned}$$

Hence, $E_{D,Y}[L_{abs}(Y, \widehat{Y})] = BE + VE + NE$, and Proposition 1 is proven. □

Meanings of absolute-error BVD

In this decomposition, bias, variance, and noise effects each has a subtractive term, and increasing the magnitude of these terms reduces rather than increases expected error, in contrast to squared-error BVD, where all characteristic terms increase expected error.

To reduce absolute-error bias effect, one can (1) reduce B and/or (2) increase the occurrence of $s_{biasEff} = -1$. Since $s_{biasEff} = -1$ when $(Y \ge \widehat{Y}$ *and* $y_o < \widehat{y}_m)$ *or* $(Y < \widehat{Y}$ *and* $y_o > \widehat{y}_m)$, it occurs more frequently when B is small (Figure 4A). That is, as absolute-error bias decreases, the fraction of bias that increases estimation error also becomes smaller (due to the increased occurrence of $s_{biasEff} = -1$). Hence, reducing model bias is double rewarded under absolute error comparing to squared error.

For variance-related terms under absolute error, both V and VE can be defined using $E_{D,Y}[L(\widehat{y}_m, \widehat{Y})|s_{varEff} = 1]P_{D,Y}$

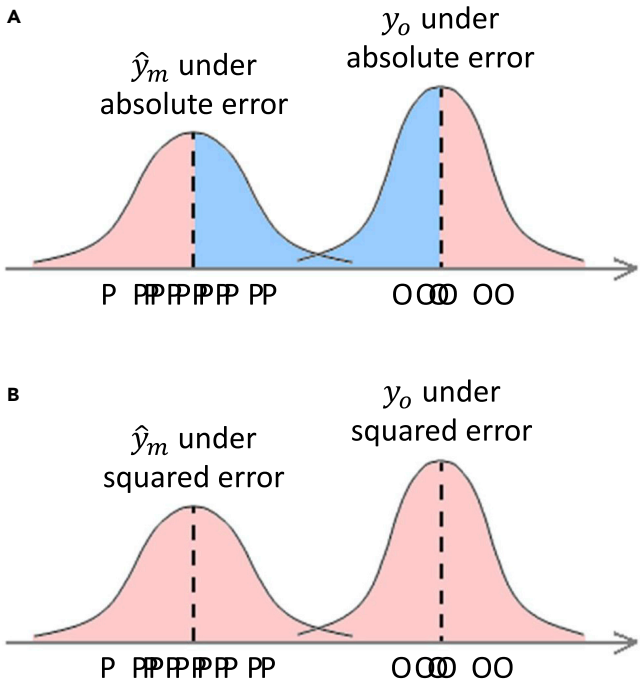

Figure 4. Different penalization patterns of absolute and squared errors

Penalization patterns of (A) absolute error and (B) squared error: red areas increase expected error, and blue areas reduce expected error.

$(s_{varEff} = 1)$ and $E_{D,Y}[L(\hat{y}_m, \hat{Y}) \mid s_{varEff} = -1]P_{D,Y}(s_{varEff} = -1)$.

If we name the former "positive variance" (i.e., the amount of variance that increases expected error) and the latter "negative variance" (i.e., the amount of variance that reduces expected error), then $V = V_{pos} + V_{neg}$ and $VE = V_{pos} - V_{neg}$. Since "variance" captures the total amount of variation, it is the sum of the amounts of both types of variances, while for "variance effect," the two types of variances each carry a sign (+/−) representing their respective effects on expected error. To reduce absolute-error variance effect, one can (1) increase V_{neg} and/or (2) decrease V_{pos}. In Figure 4A, this means to increase the blue area and decrease the red area. In contrast, optimizing squared error calls for minimizing all deviations (Figure 4B). It is known that absolute error penalizes model variation less harshly than squared error, while our BVD-based comparison illustrates exactly what types of model variations are viewed differently by the two loss functions.

Noise-related terms can be interpreted in similar ways to variance-related ones. Under absolute error, certain variations of Y (i.e., the blue area in Figure 4A) reduce expected error, making absolute error more forgiving of observational noise than squared error.

Altogether, absolute and squared errors can lead to different conclusions about the same modeling practice in model evaluation, optimization, and selection. For example, if a modeling technique reduces the blue variance in Figure 4A, it will be viewed as rewarding by squared-error-based evaluation but harmful by absolute-error-based evaluation. This perceptional difference can further result in different choices in model selection.

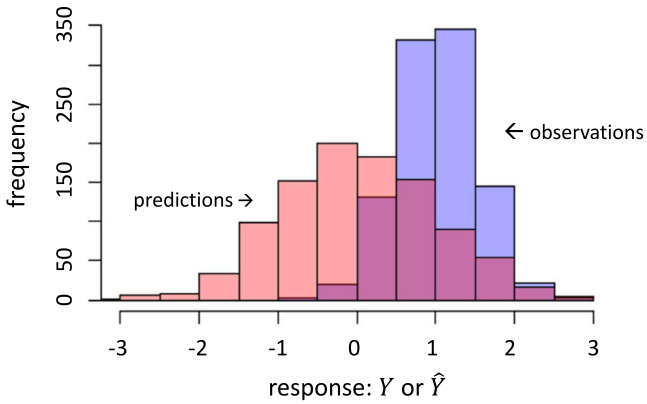

Figure 5. Histograms of the observations (shown as transparent blue bars) and the predictions (shown as transparent red bars) of the example data point

In comparison, absolute error favors low-bias models and squared error low-variance ones. As a result, absolute-error-based model selections are more likely to choose sophisticated models over simple ones, because sophisticated models tend to show lower bias, although higher variance. Meanwhile, if absolute error is the natural error metric for a specific application, using squared error for reasons like mathematical tractability or convenience is problematic, especially when evaluating sophisticated models.

The differences between the penalization patterns of the two loss functions contribute to their different behaviors in many statistical analyses. For example, in ridge regressions, squared-error-based penalty shrinks all coefficients in a relatively balanced fashion which keeps the variation among coefficients low, whereas, in lasso regressions, absolute-error-based penalty would reduce some coefficients to zero while maintaining (and sometimes increasing) others, which is analogical to reducing the red variance and preserving the blue variance in Figure 4A.

Absolute-error versus squared-error BVD for an example data point

Although normality is not required for BVD analyses, the example data point's observations and predictions (provided as Data S1) happen to be close to normal: roughly, observations $\sim N(1, 0.25)$ and predictions $\sim N(0, 1)$ (Figure 5). Two patterns emerged from applying BVD to this data point (functions and example codes provided as Data S2) (Table 1): (1) magnitudes of squared-error BVD components are larger than magnitudes of corresponding absolute-error BVD components. Not only squared error is measured in a higher dimension than absolute error, but also squared error amplifies outliers. As a result, even the square roots of squared error components are naturally larger than their absolute-error counterparts. (2) Absolute error better illustrates the relative importance of different BVD error sources than squared error. For the example data point, absolute error shows that the variance effect is about 75% of the bias effect, while squared error weighs the two effects similarly. This is because absolute error recognizes that negative variance reduces expected error, and squared error deems all variance harmful. Hence, absolute-error BVD is more effective at aiding the identification of dominant BVD error sources.

Table 1. Magnitudes of and ratios between different BVD effects for the example data point, indicating relative importance of the three BVD error sources (bias, variance, and noise), under absolute and squared errors

	B effect	*V effect*	*N effect*	$\frac{V\ effect}{B\ effect}$	$\frac{N\ effect}{B\ effect}$
Absolute error	0.62	0.47	0.12	0.75	0.19
Squared error	0.95	0.96	0.26	1.01	0.27
Square root of squared error components	0.98	0.98	0.51	1.00	0.52

Similar conclusions are found with the empirical experiments (more details below) and an in-depth case study in geospatial modeling[6]. When a model's estimation error is primarily bias in some part of the study area and primarily variance in another part, the analyst can effectively improve model performance by applying different methods (targeting bias or variance) to different parts of the study area. When subdividing the feature space, additional training data might be needed to guarantee all models for different zones of the feature space can be properly trained. When available training data are not sufficient to support the divide-and-conquer strategy, the BVD insight can help inform users of the modeling results about where and when the model's performance is trustworthy for different decision making tasks (e.g., those are sensitive to model bias versus variance).

Empirical results

In the empirical experiments, I followed the convention to assume 0 noise, because it is unattainable for the UCI datasets. The results hence consist of only bias- and variance-related components. To enable direct comparison across datasets and loss functions despite their different units, I standardized average error estimates at the dataset level. The standardized error estimates are dimensionless percentages, and the differences between various error estimates is in the unit of "percentage point" (pp).

Model complexity under absolute error: Bias/variance trade-off may not show

Although the bias/variance trade-off has been widely discussed using squared error, in our experiments, the trade-off was often absent for absolute error. This is consistent with findings of empirical works done for zero-one error[1,3], which reported erratic absence of the trade-off for empirical classification models. Together, the results suggest that the bias/variance trade-off may not be as prevalent as previous literature believed. It also gives hope to the fact that, perhaps in some practical settings, concurrently reducing bias and variance effects is possible.

In our experiments, the negative term of variance was unresponsive to complexity parameter tuning, indicating that existing modeling techniques are unable to maneuver this term. Since increasing the magnitudes of negative-effect terms can reduce expected error, new methods that treat positive and negative variances separately are needed to better minimize absolute error.

Ensembles under absolute error: Might harm model performance

Our results show that ensembles reduce the magnitudes of all variances regardless of their effects on estimation accuracy. As a result, performance gain of ensembles reported by standardized absolute error is usually smaller than standardized squared error, because, under absolute error, the benefits of reducing positive variance is partly offset by (1) the also reduced negative variance and (2) the increased ratio of bias effect to bias. Altogether, under absolute error, applying ensembles to highly biased models may increase expected modeling error, in contrast to under squared error, where ensembles never hurt model performance. Similar conclusions have been found for zero-one error[8] also due to its negative variance terms.

Absolute error versus squared error: Distinctly different perspectives

Absolute and squared errors cannot be used as proxies of each other when evaluating regression models: in our experiments, standardized squared-error estimates of modeling errors and performance adjustments were almost always larger than standardized absolute-error estimates, and sometimes by substantial amounts (up to 21 pp); for some BVD components, squared-error estimates do not resemble absolute-error estimates at all (e.g., for variance effects). The two loss functions can give different impressions about the same model or model alteration technique (e.g., absolute and squared errors view ensembles differently). Therefore, they should not be used interchangeably.

To choose an appropriate loss function is to find an error metric whose strength aligns with the objective of the analysis at hand: absolute error is more effective for reducing estimation residuals, and squared error is useful when model stability is emphasized.

The results also show that absolute-error BVD exhibits stronger contrast between different effect quantities, which makes it clearer for analysts to identify the relative importance of different BVD error sources. This feature is especially useful for delineating high-variance-effect zones versus high-bias-effect zones within the feature space and treat them with different model improvement techniques.[6]

Effects versus characteristics: Important to distinguish

The numerical differences between absolute-error effects and characteristics are notable in the empirical experiments, which reiterates the importance of distinguishing the two sets of quantities. Furthermore, the precise relationship between model characteristics and their effects, revealed by the analytically derived BVD, is useful for designing model alterations that can effectively achieve desired performance change.

In the experiments, using characteristics in place of effects (for loss functions other than squared error) showed similar consequences to using squared error in place of other loss functions, i.e., (1) overestimating the relative importance of variance effect and (2) adding ambiguity for identifying the dominant source of estimation error.

Conclusions

This research developed the first analytically derived absolute-error BVD, examined its empirical behaviors, and compared them with those of other loss functions. Generally, absolute error

is more useful for reducing estimation residuals, and squared error is better at reducing model instability. Our results show that different loss functions can have very different penalization patterns, and hence reward different model traits and respond to different model improvement strategies. Existing commonly used model improvement techniques, such as ensembles and complexity parameter tuning, were based on the characteristic-effect relationships found for squared error, which do not hold for other loss functions. As a result, applying ensembles to highly biased models under absolute error can harm expected estimation accuracy, and the widely believed bias/variance trade-off is often absent in empirical experiments under absolute error. To effectively minimize loss functions other than squared error, new methods that consider the characteristic-effect relationships under other loss functions are needed, and the analytically derived BVD provides a precise probe into such relationships. While new modeling techniques are being developed, one workaround exists, that is, to divide the feature space into zones dominated by different error sources (bias, variance, or noise) and treat the zones separately with different model alterations. For this strategy, absolute-error BVD provides clearer contrast among different error sources than squared error. It therefore may better help data-driven modelers identify dominant error sources as they vary across the feature space, and efficiently design model improvement strategies for regression models of continuous data.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Jing Gao (jinggao@udel.edu).

Materials availability

This research did not generate any materials.

Data and code availability

All datasets used in the empirical experiments are publicly downloadable from the UCI Machine Learning Repository https://archive.ics.uci.edu/.

The example data point's observations and predictions are provided as Data S1.

The functions and example codes used to conduct absolute-error and squared-error BVD analyses are provided as Data S2.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2021.100309.

ACKNOWLEDGMENTS

The author thanks Drs. Jude Shavlik and James E. Burt at the University of Wisconsin – Madison for comments on an early draft. This research was partly supported by Graduate Women in Science through a Ruth Dickie Research Fellowship.

AUTHOR CONTRIBUTIONS

J.G. conducted the research and wrote the manuscript.

DECLARATION OF INTERESTS

The author declares no competing interests.

Received: December 8, 2020
Revised: May 15, 2021
Accepted: June 18, 2021
Published: July 19, 2021

REFERENCES

1. James, G.M. (2003). Variance and bias for general loss functions. Machine Learn. *51*, 115–135. https://doi.org/10.1007/BF00058655.
2. Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. Neural Comput. *4*, 1–58. https://doi.org/10.1162/neco.1992.4.1.1.
3. Domingos, P. (2000). A unified bias-variance decomposition and its applications. In Proceedings of the 17th International Conference on Machine Learning, P. Langley, ed. (Morgan Kauffmann Publishers), pp. 231–238.
4. Willmott, C.J., and Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. *30*, 79–82. https://doi.org/10.3354/cr030079.
5. Willmott, C.J., Matsuura, K., and Robeson, S.M. (2009). Ambiguities inherent in sums-of-squares-based error statistics. Atmos. Environ. *43*, 749–752. https://doi.org/10.1016/j.atmosenv.2008.10.005.
6. Gao, J., and Burt, J.E. (2017). Per-pixel bias-variance decomposition of continuous errors in data-driven geospatial modeling: a case study in environmental remote sensing. ISPRS J. Photogramm. Remote Sens. *134*, 110–121. https://doi.org/10.1016/j.isprsjprs.2017.11.001.
7. Bartlett, P., Freund, Y., Lee, W.S., and Schapire, R.E. (1998). Boosting the margin: a new explanation for the effectiveness of voting methods. Ann. Stat. *26*, 1651–1686. https://doi.org/10.1214/aos/1024691352.
8. Breiman, L. (1996). Bagging predictors. Machine Learn. *24*, 123–140. https://doi.org/10.1007/BF00058655.
9. Breiman, L. (1996). Bias, Variance, and Arcing Classifiers (University of California, Berkeley).