Critique of model evaluation by WA Dept of Ecology

Mark Scheuerell

10/31/2019

Error in assumptions

In the section titled "Uncertainty in Dissolved Oxygen Depletion Estimates" (p59), there is an error in the assumed relationship between the standard deviation of the predictions and the root mean squared error (RMSE) of the predictions. Specifically, the document incorrectly suggests that

$$\text{Var}(\hat{y}) = \text{RMSE}(\hat{y})^2 \;\Rightarrow\; \text{SD}(\hat{y}) = \text{RMSE}(\hat{y}).$$

Specifically, the standard deviation of $\hat{y}$ is based upon differences between the predicted data points $\hat{y}_i$ and their mean $m_{\hat{y}}$

$$\text{SD}(\hat{y}) = \sqrt{\frac{(\hat{y}_i - m_{\hat{y}})^2}{N}},$$

whereas the RMSE of $\hat{y}$ is based upon differences between the predicted data points and their corresponding observed values x_i

$$\text{RMSE}(\hat{y}) = \sqrt{\frac{(\hat{y}_i - y_i)^2}{N}}.$$

Thus,

$$\sqrt{\frac{(\hat{y}_i - m_{\hat{y}})^2}{N}} \neq \sqrt{\frac{(\hat{y}_i - y_i)^2}{N}}$$
$$\text{SD}(\hat{y}) \neq \text{RMSE}(\hat{y}).$$

An example

Here is a simple example that shows how $\text{SD}(\hat{y})$ and $\text{RMSE}(\hat{y})$ are different. Consider a case where we had reason to believe that a variable y was a function of another variable x. In effort to undercover the nature of their relationship, we collected 20 samples of both y and x (Figure 1).

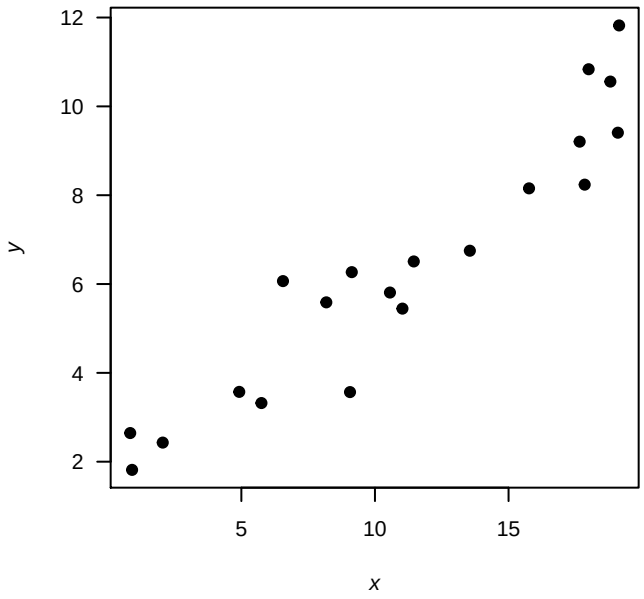

Figure 1. Plot of some hypothetical data.

Based on the apparent relationship between x and y, we might assume that each of the observed values y_i is a linear combination of an intercept β_0, the effect β_1 of a covariate x_i, and some random observation error ϵ_i, such that

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

and $\epsilon_i \sim \mathrm{N}(0, \sigma)$. We could easily estimate the unknown parameters in this model (β_0, β_1, σ), and then use the deterministic portion of the model to make predictions to compare with each of the observed values. Specifically, the predictions ($\hat{y}_i$) would be given by a straight line, such that

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

We could then use these predictions to estimate both their SD and RMSE (Figure 2). The SD of $\hat{y}$ is ~2.82, but the RMSE is only ~0.94, which is about 3 times less.

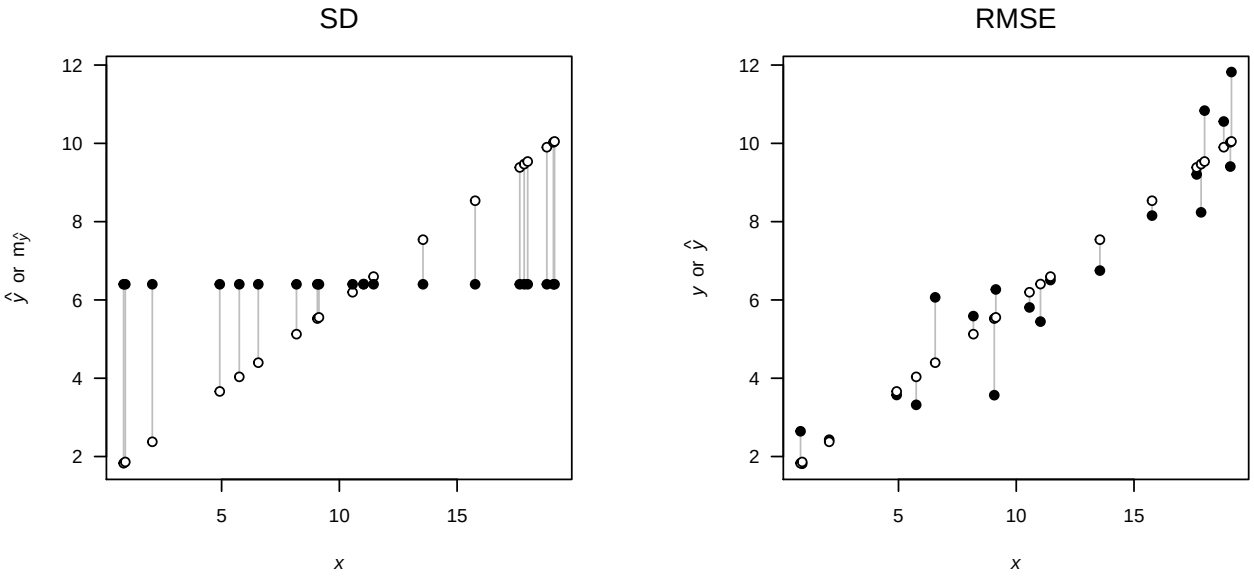

Figure 2. Graphical examples of the difference between the SD of the predictions (left) and the RMSE of the predictions (right). For the SD, the comparison is based upon the differences between the predictions (open circles) and their mean (filled circles). For the RMSE, the comparison is based upon differences between the predictions (open circles) and the observed data (filled circles). In both cases, one would square the length of each of the vertical gray lines, sum them up, and divide by the number of them before finally taking the square root.

Prediction errors

The above example dismisses an important aspect of RMSE: it should be used to compare "out of sample" predictions. Furthermore, RMSE give us an indication as to the predictive error, *on average*, rather than the uncertainty in a specific prediction.

Returning to our example above, we could estimate our uncertainty around the fitted relationship between x and y with a confidence interval (CI), which would give us an indication of the range of where the "true" fitted values would lie had we repeated our sampling exercise many times. Specifically, a $(1-\alpha)100\%$ CI on the expected relationship between x and y at some value x_k is given by

$$\hat{y}_i \pm t_{\alpha/2,n-2}\sqrt{\sigma\left(\frac{1}{n}+\frac{(x_k-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}.$$

The interval increases as the distance between x_k and $\bar{x}$ increases (Figure 3).

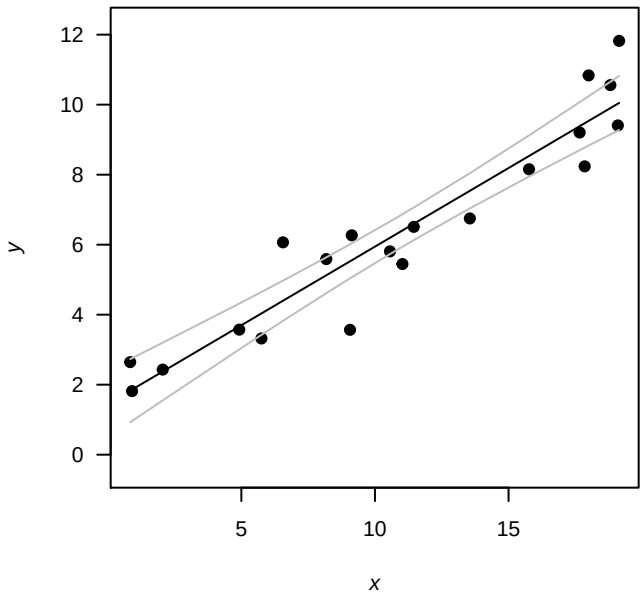

Figure 3. Confidence interval (gray lines) around the expected relationship between x and y (black line).

In a case like this, however, where we wish to make out-of-sample predictions about some new state of nature, our uncertainty around any single prediction will be necessarily greater. Specifically, a $(1-\alpha)100\%$ prediction interval (PI) around $\hat{y}$ at some value x_k is given by

$$\hat{y} \pm t_{\alpha/2,n-2}\sqrt{\sigma\left(1+\frac{1}{n}+\frac{(x_k-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}.$$

Here the paranthetic multiplier on the residual variance σ has increased by 1, which creates more uncertainty in the prediction interval than the confidence interval (Figure 4). This is because the CI only needs to account for uncertainty in estimating the expected value of y whereas the PI needs to account for a random future value of y.

So, for example, if we wanted to predict what we would observe for y if $x = 10$ with 95% certainty, we would get 5.94 ± 2.13 (Figure 5).

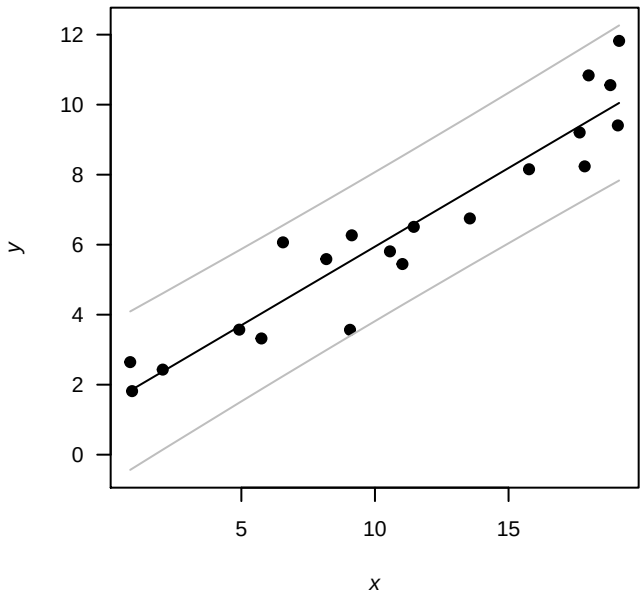

Figure 4. Prediction interval (gray lines) for future unobserved values of y.

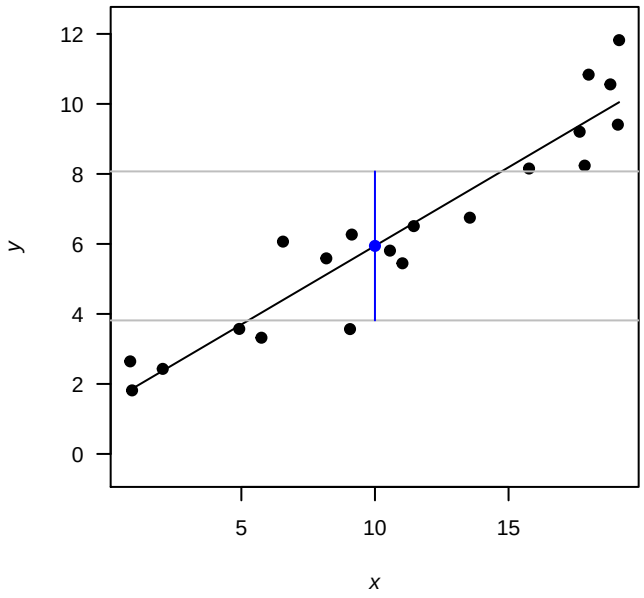

Figure 5. Example of the uncertainty around a new prediction.