Exact Standard Deviations of the Estimated Mean *Y* and Forecast Individual *Y* in Simple Regression

In a simple linear regression, the estimate of the mean of *Y* at a given x_0, as well as the forecast of an individual value of *Y* at a given x_0, are both equal to the same number – namely, the plug-in value $\hat{\alpha} + \hat{\beta}x_0$, where $\hat{\alpha}$ and $\hat{\beta}$ are the least-squares estimates of the intercept and slope. Although the values of the estimate and the forecast are the same, their uncertainties are different:

- The estimated standard deviation of the estimate of the mean of *Y* at a given x_0 is $\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$, where $\hat{\sigma}$ is the RMSE.

If the x_0 where the estimate is being made is the mean *x*, then $x_0 - \bar{x} = 0$, so $\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$ = $\hat{\sigma}\sqrt{\frac{1}{n}}$, which corresponds to the verbal rule "sigma over the square root of *n*". But this is actually just a lower bound, rather than a reliable approximation. The estimated standard deviation grows larger, the further the point of estimation x_0 is from the mean *x*. For example, if x_0 is about one *x* standard deviation from the *x* mean, then $\frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2} \approx \frac{1}{n}$, so $\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \approx \hat{\sigma}\sqrt{\frac{2}{n}}$. And if x_0 is about two *x* standard deviations from the *x* mean, then $\frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2} \approx \frac{4}{n}$, so $\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \approx \hat{\sigma}\sqrt{\frac{5}{n}}$. Although these may be small values because *n* may be large, $\hat{\sigma}\sqrt{\frac{1}{n}}$ is not a good approximation to them in a relative sense in the tails of the *x* distribution.

- The estimated standard deviation of the forecast of the value of *Y* at a given x_0 is $\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$, where $\hat{\sigma}$ is the RMSE.

This formula differs from the corresponding formula for the estimated mean (above) in having an extra 1 inside the square root. If the x_0 where the estimate is being made is the mean *x* and if *n* is large, then $x_0 - \bar{x} = 0$ and $\frac{1}{n} \approx 0$, so $\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \approx \hat{\sigma}$, which is just RMSE, the typical magnitude of individual residuals. As long as n is large, this is a good approximation. But it is actually just a lower bound. The estimated standard deviation grows larger, the further the point of estimation x_0 is from the mean x. But even if x_0 is about two x standard deviations from the x mean, then $\frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \approx \frac{4}{n}$, so $\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \approx \hat{\sigma}\sqrt{1 + \frac{5}{n}}$, which is still close to $\hat{\sigma}$ = RMSE if n is large.