Which Bootstrap When?

36-402, Spring 2013

When we bootstrap, we try to approximate the sampling distribution of some statistic (mean, median, correlation coefficient, regression coefficients, smoothing curve, difference in MSEs...) by running simulations, and calculating the statistic on the simulation. We've seen three major ways of doing this:

- The parametric bootstrap: we estimate the model, and then simulate from the estimated model;
- Resampling residuals: we estimate the model, and then simulate by resampling residuals to that estimate and adding them back to the fitted values;
- Resampling cases or whole data points: we ignore the estimated model completely in our simulation, and just re-sampling whole rows from the data frame.

Which kind of bootstrap is appropriate depends on how much trust we have in our model.

The parametric bootstrap trusts the model to be completely correct for *some* parameter value. In, e.g., regression, it trusts that we have the right shape for the regression function *and* that we have the right distribution for the noise. When we trust our model this much, we could in principle work out sampling distributions analytically; the parametric bootstrap replaces hard math with simulation.

Resampling residuals doesn't trust the model as much. In regression problems, it assumes that the model gets the *shape* of the regression function right, but doesn't make any assumption about how the residuals are distributed. It is therefore more secure than parametric bootstrap.[1]

Finally, resampling cases assumes nothing at all about either the shape of the regression function or the distribution of the noise, it just assumes that each data point (row in the data frame) is an independent observation. Because it assumes so little, and doesn't depend on any particular model being correct, it is very safe.

The reason we do not always use the safest bootstrap, which is resampling cases, is that there is, as usual, a bias-variance trade-off. Generally speaking, if we compare three sets of bootstrap confidence intervals on the same data for the same statistic, the parametric bootstrap will give the narrowest intervals, followed by resampling residuals, and resampling cases will give the loosest bounds. If the model really *is*

[1] You could also imagine simulations where we presume that the residuals take a very particular form (e.g., a t-distribution with 10 degrees of freedom), but are agnostic about the shape of the regression function, and learn that non-parametrically. It's harder to think of situations where this is really plausible, however.

correct about the shape of the curve, we can get more precise results, without any loss of accuracy, by resampling residuals rather than resampling cases. If the parametric model is also correct about the distribution of noise, we can do even better with a parametric bootstrap.

To sum up: resampling cases is safer than resampling residuals, but gives wider, weaker bounds. If you have good reason to trust a model's guess at the shape of the regression function, then resampling residuals is preferable. If you don't, or it's not a regression problem so there are no residuals, then you prefer to resample cases. The parametric bootstrap works best when the over-all model is correct, and we're just uncertain about the exact parameter values we need.