Bootstrapping Method for www.r3eda.com
14 June 2016 – R. Russell Rhinehart

Bootstrapping

This is extracted from the book, Nonlinear Regression Modeling for Engineering Applications: Modeling, Model Validation, and Enabling Design of Experiments, by R. R. Rhinehart, John Wiley & Sons, on target for release in September 2016.

One assumption in bootstrapping is that the experimental data that you have represents the entire population of all data realizations, including all nuances in relative proportion. It is not the entire possible population of infinite experimental runs, but it is a surrogate of the population. A sampling of that data then, represents what might be found in an experiment. Another assumption is that the model cannot be rejected by the data, the model expresses the underlying phenomena. In bootstrapping:

1. Sample the experimental data with replacement (retaining all data in the draw-from original set) to create a new set of data. The new set should have the same number of items in the original, but some items in the new set will likely be duplicates, and some of the original data will be missing. This represents an experimental realization from the surrogate population.
2. Using your preferred nonlinear optimization technique, determine the model coefficient values that best fit the data set realization from Step 1. This represents the model that could have been realized.
3. Record the model coefficients.
4. For independent variable values of interest, determine the modeled response. You might determine the y-value for each experimental input x-set. If the model is needed for a range of independent variable values, you might choose ten x-values within the range and calculate the model y for each.
5. Record the modeled y-values for each of the desired x-values.
6. Repeat Steps 1-5 many times (perhaps over 1,000).
7. For each x-value, create a histogram of the 1,000 (or so) model predictions. This will reflect the distribution of model prediction values due to the vagaries in the data sample realizations. The variability of the prediction will indicate the model uncertainty due to the vagaries within the data.
8. Choose a desired confidence interval value. The 95% range is commonly used.
9. Use the cumulative distribution of model predictions to estimate the confidence interval on the model prediction. If the 95% interval is desired, then the confidence interval will include 95% of the models; or, 5% of the modeled y-values will be outside of the confidence interval. As with common practice, split the too high and too low values into equal probabilities of 2.5% each, and use the 0.025 and 0.975 cdf values to determine the y-values for the confidence interval.

This bootstrapping approach presumes that the original data has enough samples covering all situations so that it represents the entire possible population of data. Then the new sets (sampled with replacement) represent legitimate realizations of sample populations. Accordingly, the distribution of model prediction values from each re-sampled set represent the distribution that would arise if the true population were independently sampled.

Bootstrapping assumes: the limited data represents the entire population of possible data, that the experimental errors are naturally distributed (there are no outliers or mistakes, not necessarily Gaussian distributed, but the distribution represents random natural influences), and that the functional form of the model matches the process mechanism. Then a random sample from your data would represent a sampling from the population; and for each realization, the model would be right.

If there are N number of original data, then sample N times with replacement. Since the Central Limit Theorem indicates that variability reduces with the square root of N, using the same number keeps the variability between the bootstrapping samples consistent with the original data. In Step 1, the assumption is that the sample still represents a possible realization of a legitimate experimental test of the same N. If you use a lower number of data in the sample, M, for instance, then you increase the variability on the model coefficient values. You could accept the central limit theorem and rescale the resulting variability by square root of M/N. But, the practice is to use the same sample size as the "population", to reflect the population uncertainty on the model.

In Step 6, if only a few re-samplings, then there are too few results to be able to claim what the variability is with certainty. As the number of Step 6 re-samplings increases the Step 9 results will asymptotically approach the representative 95% values. But, the exact value after infinite re-samplings is not the truth, because it simply reflects the features captured in the surrogate population of the original N data, which is not actually the entire population. So, balance effort with precision. Perhaps 20 re-samplings will provide consistency in the results. On the other hand, it is not unusual to have to run 100,000 trials to have Monte Carlo results converge.

One can estimate the number of re-samplings, n, needed in Step 6 for the results in Step 9 to converge from the statistics of proportions. From a binomial distribution the standard deviation on the proportion, p, is based on the proportion value and the number of data:

$$\sigma_p = \sqrt{p(1-p)/n} \tag{1}$$

Desirably, the uncertainty on the proportion will be a fraction of the proportion:

$$\sigma_p = fp \tag{2}$$

Where the desired value of f might be 0.1.

Solving Equation (17.1) for the number of data required to satisfy Equation (17.2)

$$n = \left(\frac{1}{p} - 1\right) / f^2 \tag{3}$$

If p=0.025 and f=0.1, then n≈4,000.

Although n=10,000 trials is not uncommon, and n=4,000 was just determined, I think for most engineering applications 100 re-samplings will provide an appropriate balance between computational time and precision. Alternately, you might calculate the 95% confidence limits on the y-values after each re-sampling, and stop computing new realizations when there is no meaningful progression in its value, when the confidence limits seem to be approaching a noisy steady state value.

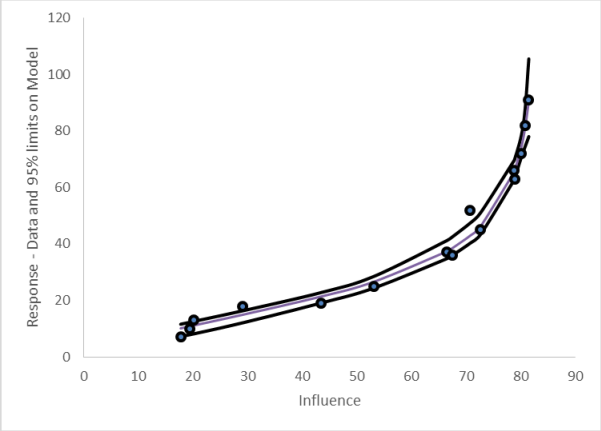
In Step 9, if you assume that the distribution of the $\tilde{y}$-predictions are normally distributed, then you could calculate the standard deviation of the $\tilde{y}$-values and use 1.96 times the standard deviation on each model prediction to estimate the 95% probable error on the model at that point due to errors in the data. Here, the term error does not mean mistake, it means random experimental normal fluctuation. The upper and lower 95% limits for the model would be the model value plus/minus the probable error. This is a parametric approach.

By contrast, searching through the n=4,000, or n=10,000 results to determine the upper and lower 97.5% and 2.5% values is a non-parametric approach. The parametric approach has the advantage that it uses values of all results to compute the standard deviation of the $\tilde{y}$-prediction realizations, and can get relatively accurate numbers with much fewer number of samples. Perhaps, n=20. However, the parametric approach presumes that the variability in $\tilde{y}$-predictions is Gaussian. It might not be. The nonparametric approach does not make assumptions about the underlying distribution, but only uses 2 samples to interpolate each of the $\tilde{y}_{0.025}$ and $\tilde{y}_{0.975}$ values. So, it requires many trials to generate truly representative confidence interval values.

Unfortunately, the model coefficient values are likely to be correlated. This means, if one value needs to be higher to best fit a data sample, then the other will have to be lower. If you plot one coefficient w.r.t. another for the 100 re-samplings and see a trend then they are correlated. When the variability on input data values are correlated, the classical methods for propagation of uncertainty are not valid. They assume no correlation in the independent variables in the propagation of uncertainty.

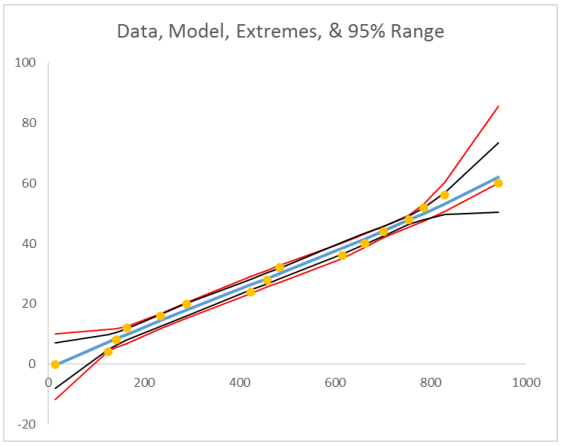
Also, Step 9 has the implicit assumption that the model matches the data, that the model cannot be rejected by the data, that the model expresses the underlying phenomena. If the model does not match the data, then bootstrapping still will provide a 95% confidence interval on the model; but you cannot expect that interval to include the 95% of the data. As a caution: If the model does not match the data (if the data rejects the model) then bootstrapping does not indicate the range about your bad model that encompasses the data, the uncertainty of your model predicting the true values.

This figure reveals the results of a Bootstrapping analysis on a model representing pilot-scale fluid flow data. The circles represent data, the inside thin line is the modeled value, and the darker lines indicated the 95% limits of the model, based on 100 realizations evaluated at the x-values of the data set of 15 elements. (The y-axis is the valve position and the x-axis is the desired flow rate. Regression was seeking the model inverse.)



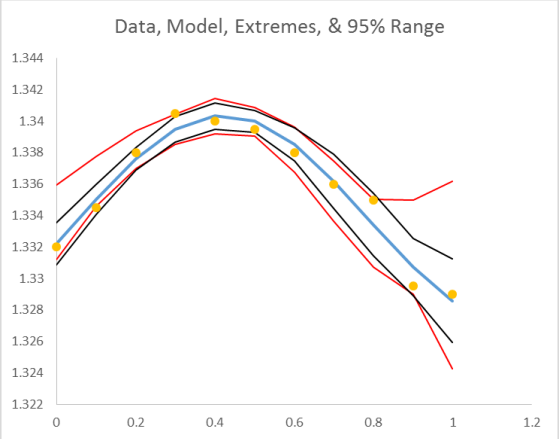
Result Interpretation

Here is a bootstrapping analysis of the response of electrical conductivity to salt concentration in water. The data are generated with [salt] as the independent variable, and conductivity as the response, but as an instrument to use conductivity to report [salt] the x and y axes are switched. Here is a calibration graph. The model is a polynomial of order 3 (a cubic power series with four coefficients).



The conductivity measurement results in a composition error of about +/- 2 mg/dL in the intermediate values and higher at the extremes. The insight is, if +/- 2 mg/dL is an acceptable uncertainty, then the calibration is good in the intermediate ranges. If not, the experimenters need to take more data to use more points to average out variation. Notice that the uncertainty in concentration has about a +/- 5 mg/dL value in the extreme low or high values. So, perhaps the experiments need to be controlled so that concentrations are not in the extreme low or high values where there is high uncertainty.

Here is another example of data from calibrating index of refraction with respect to mole fraction of Methanol in water. Because the response gives two mole fraction values for one I.R. value, the data represents the original x-y order, not the inverse that is normally used to solve for x given the measurement. This model has a sine functionality, because it gives a better fit than a cubic polynomial.



How to interpret? If the distillation group takes a sample, measures index of refraction, and gets a value of 1.336, the blue model indicates that the high methanol composition is about 0.72 mole fraction. But, the 95% limits indicate that it might range from about 0.65 to 0.80 mole fraction. This large range indicates that the device calibration produces +/- 0.07 mole fraction uncertainty, or +/- 10% uncertainty in the data. However, if the index of refraction measurement is 1.339, the model indicates about 0.4 mole fraction, but the uncertainty is from 0.3 to 0.6. Probably, such uncertainty in the mole fraction would be considered too large to be of use in fitting a distillation model to experimental composition data.

Bootstrapping analysis reveals such uncertainty, when it might not be recognized from the data or the nominal model. The model can be substantially improved with a greater number of data in the calibration.

I greatly appreciate Reed Bastie, Andrea Fenton, and Thomas Lick (fall 2015) for sharing their index of refraction data for my algorithm testing. And to Carol Abraham, John Hiett, and Emma Orth for sharing their conductivity calibration data. The fact that this analysis suggests that they need more data is not a criticism of what they did. One cannot see how much data is needed until after performing an uncertainty analysis.

Hopefully, these examples help you understand how to use and interpret model uncertainty.

Appropriating Bootstrapping

For steady-state models with multivariable inputs (MISO), or for time-dependent models, the Bootstrapping concept is the same. But, the implementation is more complex.

For convenience, I would suggest to appropriate the Bootstrapping technique. Hand select a portion of data with "+" residual values and replace them with sets that have "-" residual values. Find the model. Then replace some of the sets which have "-" residuals with data that has "+" residuals. Find the model. Accept that the two models represent the reasonable range that might have happened if you repeated the trials many times. I think this is as reasonable a compromise to Bootstrapping as Bootstrapping is to what would happen if truly generating many independent sets of trial data.

A first impression might be to take all of the data with "+" residuals out, replacing them with all of the data with "-" residuals. But don't. Consider how many Heads you expect when flipping a fair coin 10 times. You expect 5 Heads and 5 Tails. You would not claim that something is inconsistent if there were 6 and 4, or 4 and 6. Even 7 and 3, or 3 and 7 seem very probable. But 10 and 0, or 0 and 10 would be improbable, an event that is inconsistent with reality. So, don't replace all of the "+" data with "-" data or vice versa. Take a number that seems to be at the 95% probability limits. The binomial distribution can be a good guide. If the probability of an event is 50%, then with N=10 data, expecting 5 to have "+" residuals the 98% limits are about 8 and 2, or 2 and 8. So, move 3 from one category to the other. For

N=15 data the 95% limit is about 11 and 4, or 4 and 11; so, move 4. For N=20 data, it is about 14 and 6, or 6 and 14; so, move 4. Use this as a guide to choose the number of data to shift.