The Adequate Bootstrap — A new Method for Measuring Model Uncertainty

T. Kenney[1*] and H. Gu[1]

[1] *Department of Mathematics and Statistics, Dalhousie University: tkenney@mathstat.dal.ca, hgu@dal.ca*
[*] *Presenting author*

Keywords. *Model Adequacy; Robust Inference; Bootstrap.*

Model adequacy testing is less ubiquitous than it ought to be. Any parametric analysis should be acompanied by model adequacy testing. However, in practice, this is not always the case. There are several reasons for this. One particular reason is the fundamental disconnect between what is tested, and what we would like to test. The usual approach to testing model adequacy is to set up an hypothesis test. The null hypothesis is "Model M is the true model." However, when we consider the famous words of **?**: *"All models are wrong. Some models are useful."* we see the problem with this approach. We already know that the null hypothesis is false, and our model is wrong. What we want to know from our test is whether the model is useful. A model might still be useful even if we have enough data to reject it.

We consider the context where we are confident that our model reflects some part of the underlying process, but some further process (such as data contamination or sampling bias) results in observed data that do not follow the model distribution. The question we ask ourselves is how much uncertainty in our parameter estimates is caused by the difference between the model distribution and the actual data distribution.

Our solution to this problem is to use bootstrap inference on samples of a smaller size, for which the model cannot be rejected. We use the model adequacy test to choose a bootstrap size with limited probability of rejecting the model (we use probability 0.5 for analytical convenience). The intuitive idea is that if we have a sample size for which the model adequacy test is not often rejected, and our inference at this sample size gives a certain confidence interval, then we should be happy with this inference, because we might have been confident in it if our original dataset had been this size.

This approach has parallels with the *credibility index* of **?**, which uses subsampling and a model adequacy test to measure the extent to which the model matches the

data. However, the credibility index is simply a measure of how much data is needed to falsify the model. It does not give such an easily interpretable assessment of the goodness of fit, in terms of its effect on parameter estimates. That is, merely knowing that about 2,000 data points is sufficient to falsify a given model does not give a clear impression of whether the model is useful — in some cases this makes the model useful, and in others it does not. A confidence interval incorporating uncertainty due to model misspecification is often much easier to relate to usefulness.

We demonstrate the theory and application of the adequate bootstrap in two common situations — contamination and sampling bias. In both of these situations, we show that the adequate bootstrap greatly improves our coverage under the misspecified model cases. Meanwhile, when the model is not misspecified, the adequate bootstrap is able to recover the same confidence interval as inference based on the full data.

References

G. E. P. Box (1976), Science and Statistics, *Journal of the American Statistical Association* **71** 791–799

B. Lindsay and J. Liu (2009) Model Assessment Tools for a Model False World *Statistical Science,* **24**, 303–318