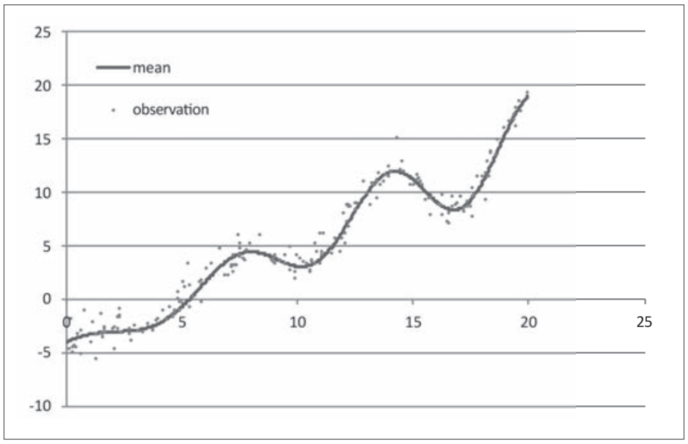
Goldilocks and the Three Modelers

By Brian Holland

It is no news that computing power is now cheap. However, this has created a new problem: when to stop modeling? It is easy to try out many different models. We intuit that it is a bad idea to have overly complex or simple models. They should be just right. But what does that mean, and how do we tell? This brief note is a reflection on an issue that we know we have, and an attempt to socialize some concepts and terms to describe it that will help the actuarial community address the issue in forming opinions about assumptions.

A construed example will help to illustrate the issue (Figure 1). We want to find the mean of the points: the black line is the underlying mean. The points could be functions of observed decrements or something else, it hardly matters for an example.

Figure 1
Construed Example: Y = 0.1 x sin x + 3 exp(-x) cos x + 0.1 Z



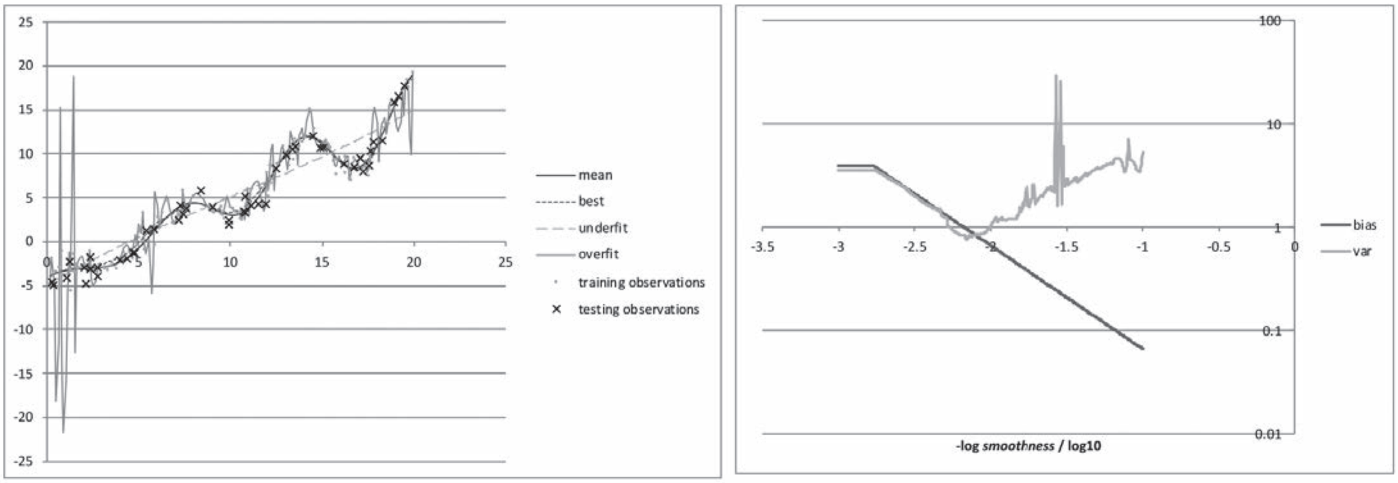
What will make a model "just right" is whether it predicts better than alternatives. To find one, we can take most of the observations, train a model on those, and test the model on the remaining samples. We can compare between different classes of models, or tune models within one class by means of a "metaparameter" that is used to adjust the level of complexity. Two examples follow: spline regression and a tree model. The important point is that whatever type of model we try, we have this issue of model complexity: choosing which features to use in varying mortality slope, or something more elaborate.

A CONTINUUM OF OPTIONS: SPLINE MODEL

The spline regression model[1] is especially nice for this sort of comparison. The metaparameter in this case is used to specify the smoothness of the curve. The higher the smoothness is, the lower the wiggliness, and vice versa. Figure 2 shows three spline regressions versus the data: one underfit, one overfit and one in between. The underfit curve is only a straight line. Smoothness is high, so wiggliness is low. The overfit curve hits almost all the training points—but that is hardly good! That means it is too complex, and will not fit the new data well. It overfits the training data. Goldilocks might say that one bed is too smooth and hard, while one is too soft and lumpy. We can compare those models on a continuum, shown on the right. The mean squared error (MSE) on the training data is the bias, while testing MSE is the variance. The underfit model uses the parameter on the left side. Note that the metaparameter is transformed so the simple model is on the left and the complexity increases to the right—that is the customary presentation. There is a trade-off between model complexity and predictive value. This trade-off is called the bias-variance trade-off. Some additional complexity helps, but after a point, it hurts the predictive value of the model. The best fit model shown on the left is the one at the minimum variance. In the bias-variance trade-off, additional complexity reduces the bias, or error on the training set.

In this case you can see some static due to numerical precision issues in the variance graph—it would ideally decrease and then increase fairly smoothly. I've left those blemishes in this presentation because you are also likely to see that sort of thing in your own experiments.

Figure 2
Spline Regression



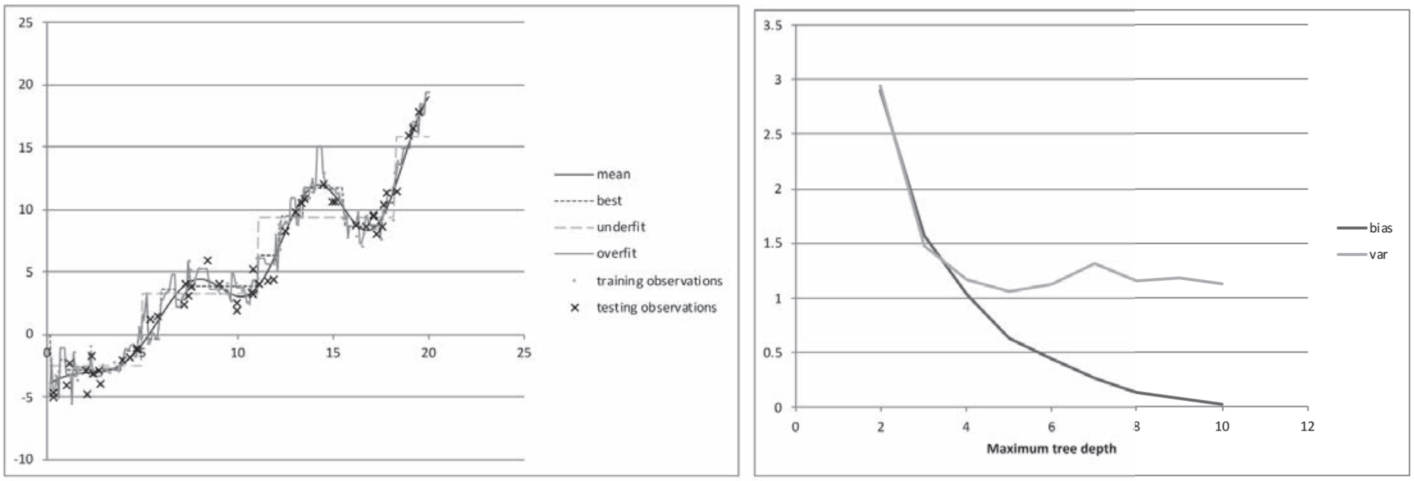
ANOTHER EXAMPLE: TREE MODEL

As actuaries recommending assumptions, we need to remain cognizant of the bias-variance trade-off and its implications for the predictive value of assumptions. This issue is present regardless of the model type. Figure 3 shows the same trade-off for a tree model. A tree model is used to split the data at a certain point to minimize error on each side of that point, using the simple mean on each side as the predictor. Then, if an additional split reduces the error further, the tree will split again. In this example the metaparameter is the maximum depth, or number of splits, that the tree is allowed to make. The underfit case is allowed at most two splits, so 4 = 2*2 averages for sections of the line. There is clearly more going on in the data and more splits will help predictions. Allowing up to 10 splits allows the model to reach out and grab the outliers, which would be poor predictions for neighboring points. The optimum value in this case is about five splits at most, i.e., 32 segments of the line. There are only 150 training points out of the 200 total, so a tree depth of seven, which allows 128 segments, allows a local space around nearly each training point and the deeper models are about the same.

COMMUNICATION

Actuaries have another job besides forecasting: communicating their decisions. That job can be at least as important. I would assert that using such demonstrations as the bias-variance trade-off will help actuaries show why they have chosen a particular degree of complexity. That justification is especially important given that a degree of art and judgment will remain in our work.

Figure 3
Decision Tree



The responsibility to communicate points us to another trade-off: not just model complexity, but communication complexity. It can take some practice to be able to wing a pithy explanation of model complexity to professionals from other spaces, such as accountants. Using a simple average or linear regression is simple to communicate, but a penalized spline model is less so—especially if discussing the choice of penalty. In this context I like to think of complexity as length of the story. A simple story is a short story, a more complex story is longer. Depending on the assignment, a simpler story might be better than a more (quantitatively) predictive but longer story. Various complexities will come up in absorbing these techniques in organizations where the techniques are unfamiliar, possibly starting with the model results if a different random subset of training data is chosen. The organizational learning curve can be long. I find it helpful to remember that we are only predicting anyway; we're just trying out how well those predictions work in advance.

Among practitioners, summary statistics can certainly facilitate communication: cross-validation statistics, AIC, or BIC, for example, get at the same underlying issue of model complexity versus predictive value.

DÉJÀ VU ALL OVER AGAIN

By now you might be wondering what seems so familiar about this issue. The model complexity trade-off is nothing new to actuaries. Whittaker-Henderson type B includes two components in its objective function, combined with a weight: a fidelity, or fit, component describing model error, and a smoothness parameter. Henderson published this approach in 1923. It was computationally expensive. These days, of course, computation should not be an issue. For a nice discussion and comparison to current methods, please see "Back to the Future with Whittaker Smoothing" by Iain Currie, *https://www.longevitas.co.uk/site/informationmatrix/whittaker.html*.

The bias-variance trade-off even appears in *Transactions of the Society of Actuaries*, 1995, Vol. 47 in "Graduation" by Kernel and "Adaptive Kernel Methods With a Boundary Correction" (Gavin, Haberman and Verrall). So, it is nothing intrinsically new in our space and is certainly fair game.

CONCLUSION

Actuaries continually face choices in assumption complexity. The conceptual framework provided by the bias-variance trade-off can help actuaries communicate their choices between overfit and underfit in their search for a model that is just right. ■

Brian D. Holland, FSA, MAAA, is director and actuary, Individual Life and A&H Experience Studies at AIG in Atlanta. He can be reached at *brian.holland@aig.com.*

ENDNOTE

[1] *https://docs.scipy.org/doc/scipy/reference/generated/scipy.interpolate.splrep.html*, please try it out yourself.