Bookmark this page

5.3.R1

0/1 point (graded)

Suppose that we perform forward stepwise regression and use cross-validation to choose the best model size.

Using the full data set to choose the sequence of models is the WRONG way to do cross-validation (we need to redo the model selection step within each training fold). If we do cross-validation the WRONG way, which of the following is true?

- The selected model will probably be too complex ✔
- The selected model will probably be too simple ✖

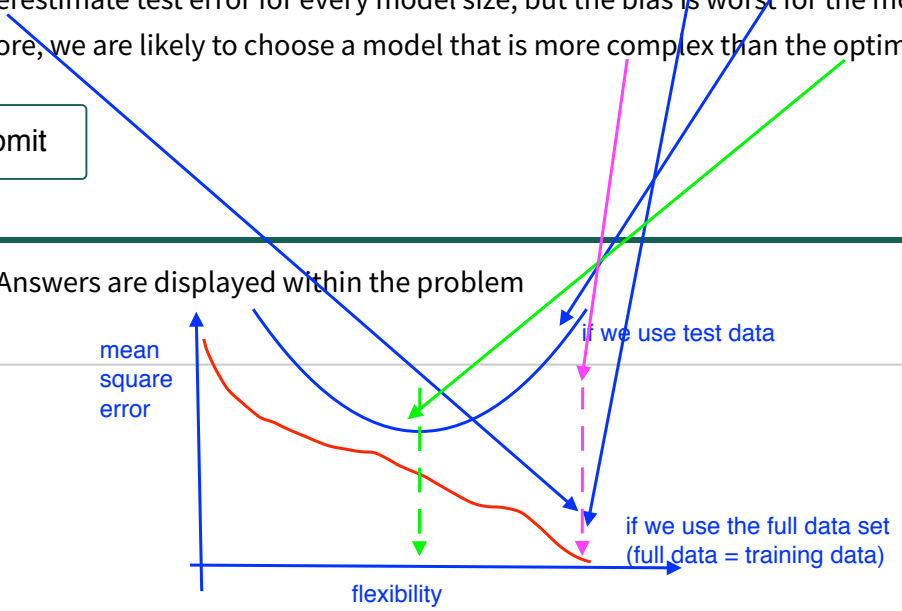
Explanation

when the model is overfitting, the price we should pay (MSE) should be high, but since we use full data set (as training set), the price we pay (MES) is small

Using the full data set to choose the best variables means that we do not pay as much price as we should for overfitting (since we are fitting to the test and training set simultaneously). This will lead us to underestimate test error for every model size, but the bias is worst for the most complex models. Therefore, we are likely to choose a model that is more complex than the optimal model.

Submit

Answers are displayed within the problem



https://stats.stackexchange.com/questions/221382/model-complexity-in-cross-validated-stepwise-regression/221494#221494

In this hypothetical wrong procedure, it sounds like the number of regressors is chosen to minimize the error over the whole data set. In this case, the model will engorge itself on regressors until using all of them, because the error will always decrease as more are added. The reason is that the error is evaluated using the same data used to choose the weights. This allows the model to overfit (i.e. to fit random structure in the training data that isn't representative of the underlying distribution that produced it). This means that the error will be optimistically biased; when run on new data from the same distribution, the model will have greater error, and will regret its former gluttony.

Regarding answer #1, 'too complex' means more regressors are chosen than should have been, leading to overfitting. This assumes that the 'proper' model includes a smaller subset of the regressors.

That said, using stepwise regression is generally not a good idea in the first place (e.g. see here).