5.3 Review Questions

5.3.R1

1/1 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?

- (selected) The selected model will probably be too complex
- The selected model will probably be too simple

✔

Explanation
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