In class exercise week 11 solution
Topic: prediction models

Problem 1 (Regression to the mean)

Read the first two parts of the article "Three things that every medical writer should know about statistics" by Stephan Senn. Then, read the description of the following study and answer the questions:

A randomized controlled trial (RCT) was conducted in order to examine if praise or criticism is helpful for medical students when learning to take blood samples from patients. Therefore, students were randomly assigned to one of two groups. In the first group, students who performed bad in taking blood samples were criticized. In group 2, teachers praised students performing very well. Next time, when the students were taking blood samples from the patients, the following observations were made: On average, the criticized students from group 1 improved while the praised students from group two worsened.

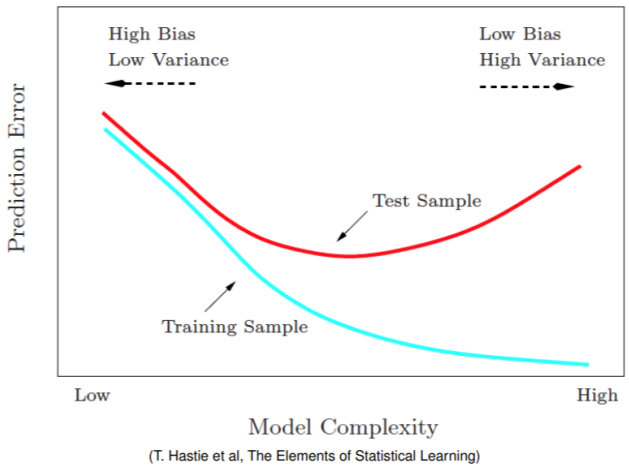
a) Based on this study, can you conclude that criticizing is beneficial to praising students when learning new task like taking blood?

   Solution: No. The observation can be explained by regression to the mean, assuming that neither praise nor criticism is helpful.

b) How could you design the study when you want to conclude that praise respectively criticism is helpful?

   Solution: Both groups would need a control group. For instance, in group 1, a subset of the bad performing students should be criticized, the other part shouldn't. Then, if the criticized students in group 1 improve, this is due to criticism. The same holds for group two.

Problem 2 (Bias Variance Tradeoff)



(T. Hastie et al, The Elements of Statistical Learning)

a) What indicates that models with high complexity lead to overfitting?

Solution: There is a large gap between the curve of the training and the test sample. The curves indicate that models with high complexity have a very low prediction error on the training but a high error on the test data. This means, that the model is so complex, that it fits the training data perfectly but it generalizes bad on new data, i.e. it overfits.

b) Based on which criterion can you find the optimal model complexity?

Solution: The model complexity should be chosen such that the test error is minimal which simply equals the minimum of the red test curve.

c) Why are low complexity models correlated with high bias?

Solution: A small complexity means that we fit a model that can capture a simple association between predictors and outcome (for example a linear model). However, this model might not be complex enough to model the real underlying relationship leading to systematic deviations.