Model selection

Types of research questions

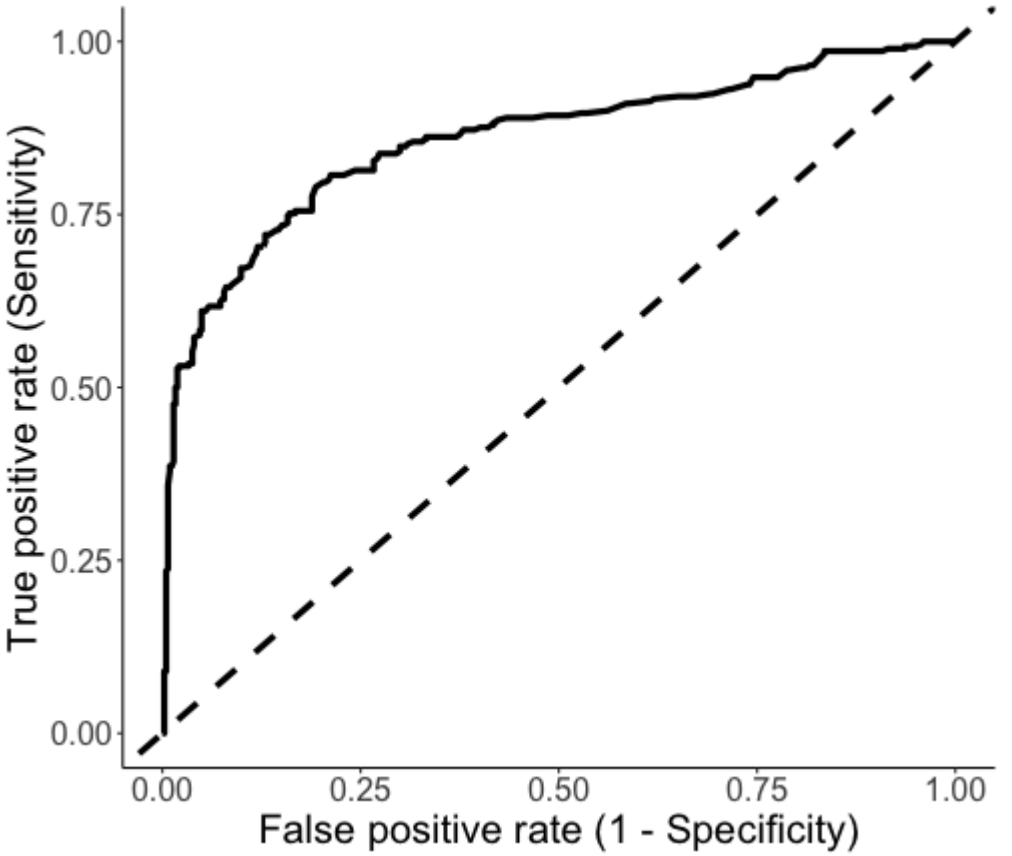
So far, we have learned how to answer the following questions:

- What is the relationship between the explanatory variable(s) and the response?
- What is a "reasonable range" for a parameter in this relationship?
- Do we have strong evidence for a relationship between these variables?
- How *well* does our model predict the response?

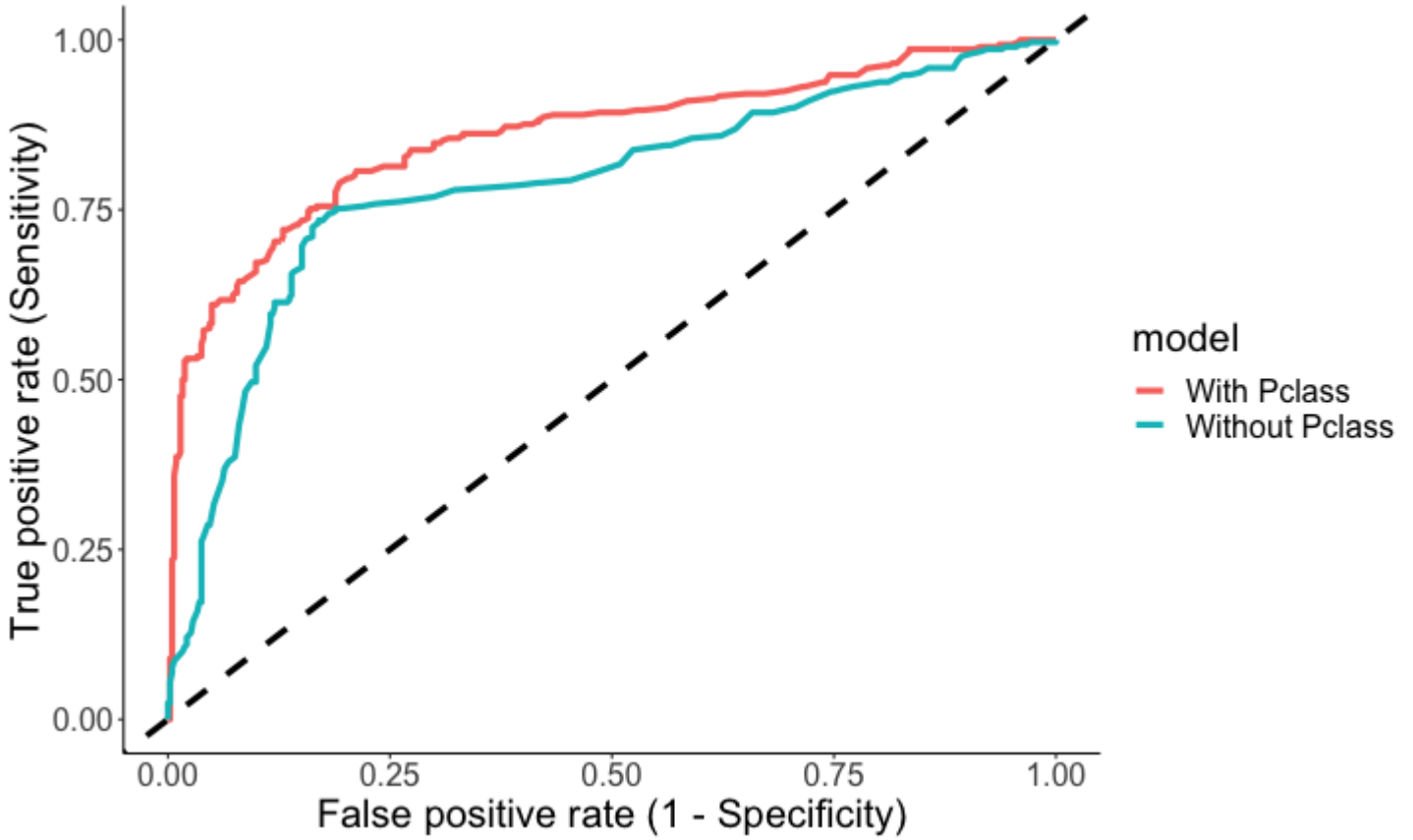
Today we will ask:

- How do we select a model when there are many possible explanatory variables?

Last time: ROC curve



Comparing models with ROC curves



Problem: reusing data...

> It is generally a bad idea to assess performance of a model on the same data we used to train it. This can lead to overfitting.
>
> What can we do instead?

Systematically comparing models

We want to select the model which best predicts the response.

When, and when not, to use model selection