Model selection

- Exam 1 Due Friday
- Project 1 Due next wednesday
- No class on Friday (project work time)

Types of research questions

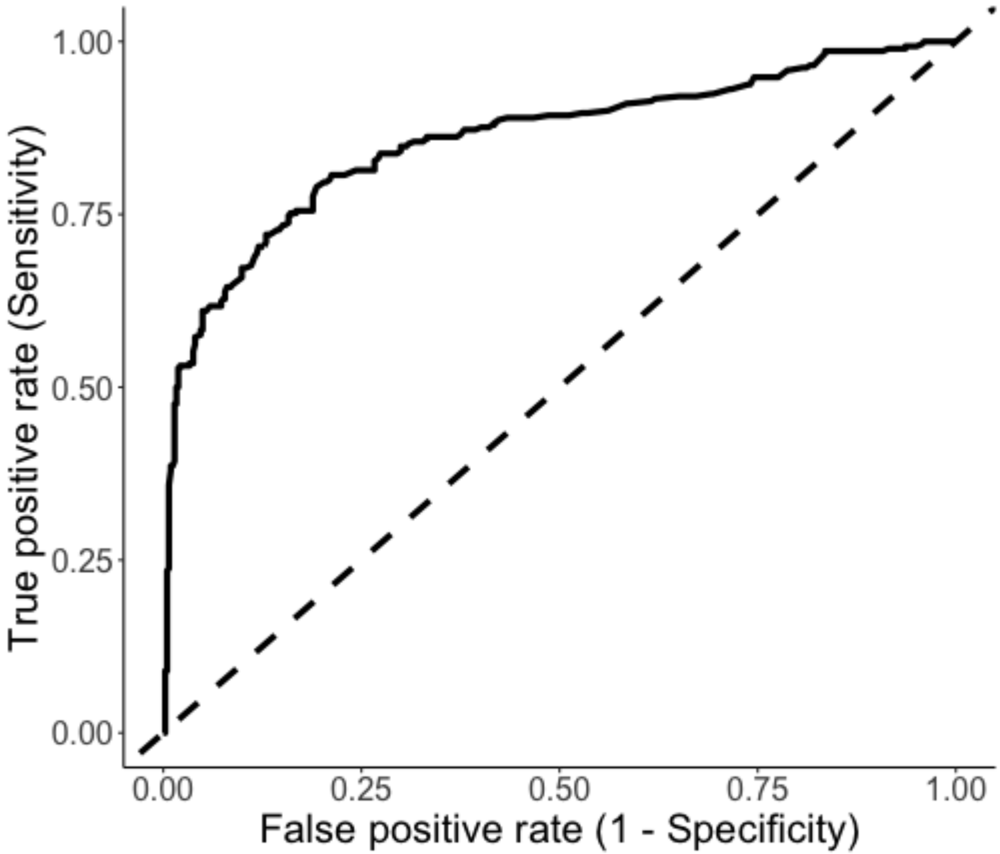
So far, we have learned how to answer the following questions:

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

Today we will ask:

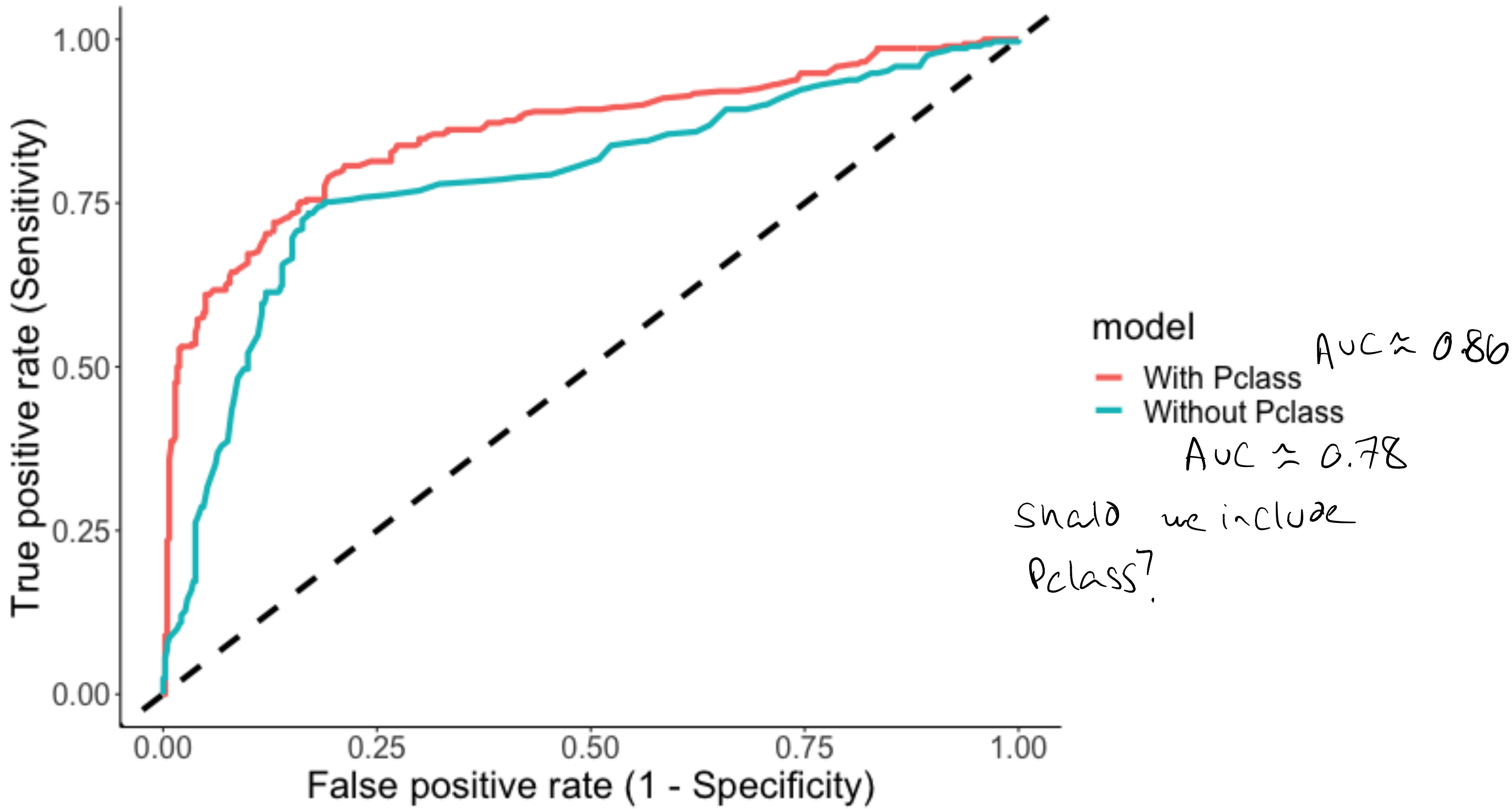
- How do we select a model when there are many possible explanatory variables? model selection!

Last time: ROC curve



AUC ≈ 0.86

Comparing models with ROC curves



Should we include Pclass?

Problem: adding more predictors should generally increase performance on training data

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?

- Divide data into training & test sets. Evaluate performance on test set
- cross validation: divide data into K "folds" For each fold, train on the other K-1 folds, evaluate on the held out fold. Average performance across K folds
- with nested models, we can do a hypothesis test
- Use another metric to compare models

comparing models based on likelihood:

AIC

with Pclass: $-2\ell(\hat{\beta}) = 635.39$ $\quad 635.39 + 2(6) = 647.39$

w/out Pclass: $-2\ell(\hat{\beta}) = 740.40$ $\quad 740.40 + 2(4) = 748.4$

when we fit the model, want to minimize $-2\ell(\hat{\beta})$

But deviance always decreases when we add parameters

Analogue (linear regression): SSE ($\downarrow$ when we add parameters)

$\Rightarrow R^2 = 1 - \frac{SSE}{SSTotal}$ $\uparrow$ when we add parameters

$R^2_{adj} = 1 - \frac{SSE / (n-(k+1))}{SSTotal/(n-1)}$ adding a penalty for # of parameters

why this penalty term?

If the model is correct,

$\ell(\hat{\beta}) - (k+1)$

is an unbiased estimate of the expected log likelihood on a new sample of data

$$AIC = \underbrace{-2\ell(\hat{\beta})}_{\text{want this small}} + \underbrace{2(k+1)}_{\text{don't want too many parameters (penalty term)}}$$

<u>Alternative</u>: BIC (Bayesian information criterion)

$$BIC = -2\ell(\hat{\beta}) + \underbrace{(k+1)\log n}_{\text{stronger penalty} \Rightarrow \text{smaller model}} \qquad (n = \#\text{obs.})$$

- AIC is similar to LOOCV (think of AIC as a fast approximation to LOOCV)

- As $n \to \infty$, we generally expect the prediction performance of a model chosen w/ AIC / LOOCV to be close to the best performing model

- As $n \to \infty$, if the <u>true</u> model is one being compared, AIC / LOOCV will tend to pick a <u>strictly larger</u> model than the truth

- As $n \to \infty$, if the <u>true</u> model is one being compared, BIC will tend to pick the true model

- Models selected by BIC tend to predict less well than models selected by AIC / LOOCV

Systematically comparing models

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

Need:

1) Method for comparing models (usually AIC or BIC)
2) Method for searching through different models

Model search algorithms:

1) Best subset selection: consider all possible models (all possible combinations of the variables)
2) Forward (stepwise) selection:
   - Start w/ "minimal" model (usually intercept-only)
   - Add terms until model stops improving

greedy algorithms

3) Backward (stepwise) selection:
   - Start w/ all possible terms in the model
   - Remove terms until model stops improving

When, and when not, to use model selection