Lecture 2

Dengue paper recap

What is the main goal of the research study?

Dengue paper recap

What model do the researchers use to predict dengue status?

Dengue paper recap

How did the researchers choose their final model?

Dengue paper recap

How did the researchers assess the performance of their model?

Types of research 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 can we predict the response / new observations?
- What model should we use to predict the response / which variables are most important?

Our next steps

- Assessing binary predictions
- Model selection
- Choosing an analysis method and designing a statistical analysis plan

Titanic data

Recall the Titanic data from last semester:

- Data on 891 passengers
- Variables include:
  - Survival
  - Sex
  - Age
  - Passenger class

Modeling Titanic data

Suppose we fit the following model:

$$\text{Survived}_i \sim \text{Bernoulli}(p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{Male}_i + \beta_2 \text{Age}_i + \beta_3 \text{Class2}_i + \beta_4 \text{Class1}_i$$

How should we assess predictive ability of the model?

Making binary predictions

- For each passenger, we calculate $\hat{p}_i$ (estimated probability of survival)
- But, we want to predict *which* passengers actually survive

Question: How do we turn $\hat{p}_i$ into a binary prediction of survival / no survival?

Confusion matrix

```r
m1 <- glm(Survived ~ Sex + Age + Pclass, data = titanic,
          family = binomial)

table("Predicted" = ifelse(m1$fitted.values > 0.5, 1, 0),
      "Observed" = m1$y)
```

```
          Observed
Predicted   0   1
        0 356  83
        1  68 207
```

Question: Did we do a good job at predicting survival?

Why a threshold of 0.5?

Another confusion matrix

Researchers fit a model for the dengue data and produce the following confusion matrix:

		Observed	
		$Y = 0$	$Y = 1$
Predicted	$\widehat{Y} = 0$	3957	1631
	$\widehat{Y} = 1$	66	66

The accuracy is 70%. Is the model doing a good job?

Changing the threshold

Threshold of 0.3:

```r
table("Predicted" = ifelse(m1$fitted.values > 0.3, 1, 0),
      "Observed" = m1$y)
```

```
          Observed
Predicted   0   1
        0 301  46
        1 123 244
```

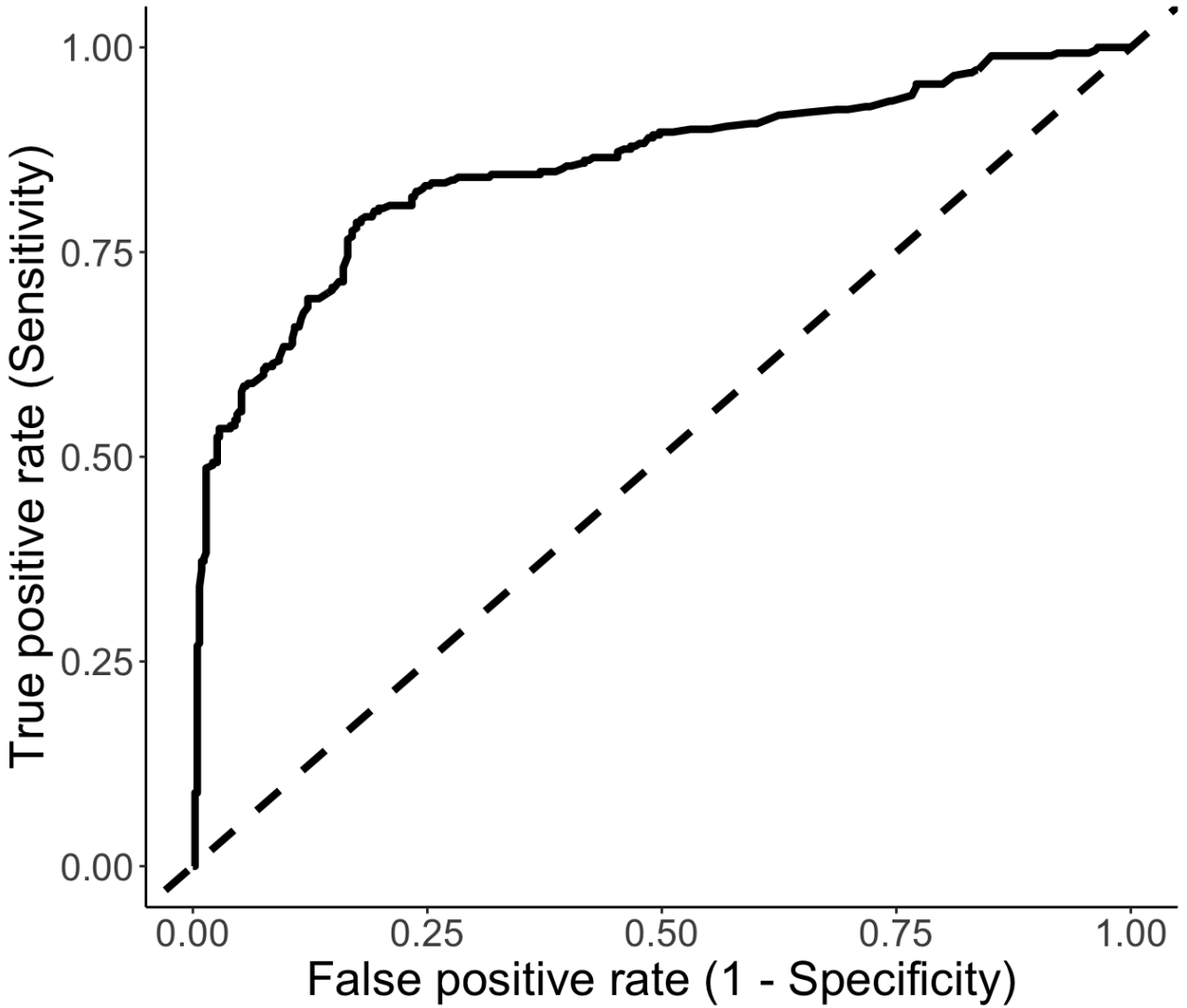
Threshold of 0.7:

```r
table("Predicted" = ifelse(m1$fitted.values > 0.7, 1, 0),
      "Observed" = m1$y)
```

```
          Observed
Predicted   0   1
        0 407 134
        1  17 156
```

How do sensitivity and specificity change?

ROC curve


1.00
0.75
0.50
0.25
0.00
True positive rate (Sensitivity)
0.00
0.25
0.50
0.75
1.00
False positive rate (1 - Specificity)

Area under the curve (AUC)

Summary

- Threshold predicted probabilities to get binary predictions
- Performance metrics like accuracy, sensitivity, and specificity can be calculated from a confusion matrix
- A threshold of 0.5 maximizes accuracy (in the population)
- As threshold increases, sensitivity decreases and specificity increases
- ROC curves plot the trade-off between sensitivity and specificity