

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?

Early Dengue Classifier: logistic regression
w/ Age, WBC, PLT

Dengue paper recap

How did the researchers choose their final model?

- model selection using AIC to find important variables
- compare predictive performance between smaller and larger models

Dengue paper recap

How did the researchers assess the performance of their model?

- sensitivity & specificity
- ROC curves, AUC

Types of research questions

- What is the relationship between the explanatory variable(s) and the response? (fitting & interpreting model)
- What is a “reasonable range” for a parameter in this relationship? (confidence interval)
- Do we have strong evidence for a relationship between these variables? (hypothesis testing)
- How well can we predict the response / new observations?
- What model should we use to predict the response / which variables are most important?

next steps

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?

compare predicted survival (\hat{y}_i) to
observed survival (y_i)

\Rightarrow need to convert \hat{p}_i into binary predictions
 $\hat{y}_i \in \{0, 1\}$

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?

$$\hat{y}_i = \begin{cases} 1 & \hat{p}_i \geq \text{threshold} \quad (\text{e.g. } 0.5) \\ 0 & \hat{p}_i < \text{threshold} \end{cases}$$

$$= \mathbb{1} \{ \hat{p}_i \geq \text{threshold} \}$$

Confusion matrix

```
1 m1 <- glm(Survived ~ Sex + Age + Pclass, data = titanic,
2           family = binomial)
3
4 table("Predicted" = ifelse(m1$fitted.values > 0.5, 1, 0),
5       "Observed" = m1$y)
```

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

Accuracy: if I randomly select an observation, what is the probability my prediction is correct?

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

True negatives

	$\hat{y} = 0$	$\hat{y} = 1$	
$\hat{y} = 0$	356	83	False negatives
$\hat{y} = 1$	68	207	Correct (True positives)

Accuracy: $\frac{\# \text{ correct predictions}}{\# \text{ observations}} = \frac{TP + TN}{n} = 0.79$

1 - Accuracy = classification error

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	$\hat{Y} = 0$	3957	1631
	$\hat{Y} = 1$	66	66

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

Changing the threshold

Threshold of 0.3:

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

	Observed	
Predicted	0	1
0	301	46
1	123	244

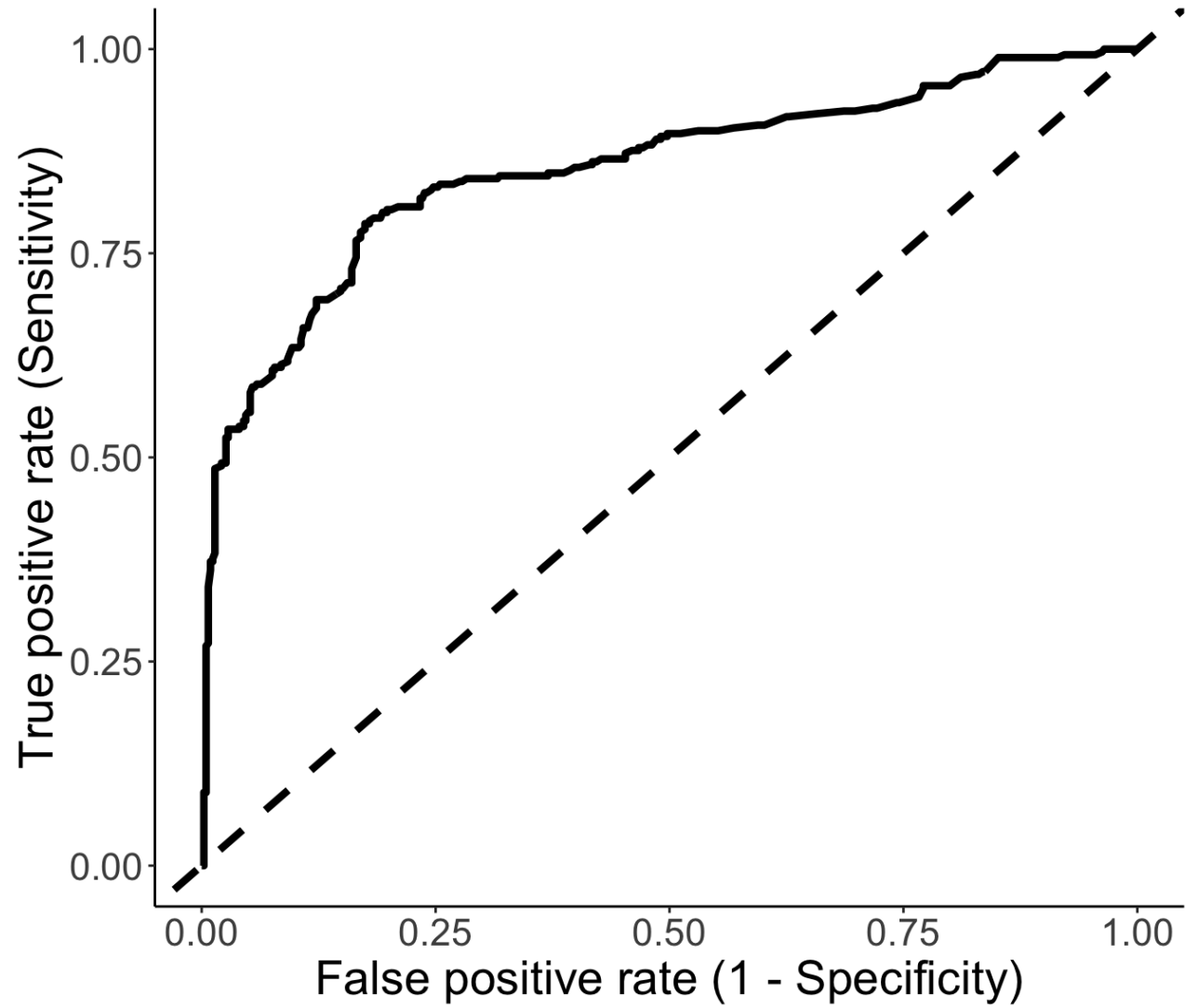
Threshold of 0.7:

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

	Observed	
Predicted	0	1
0	407	134
1	17	156

How do sensitivity and specificity change?

ROC curve



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

