Prediction

Types of research questions

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

- What is the relationship between the explanatory variables and the response? fit a 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 do we select a model when there are many possible explanatory variables? variable

Today we will ask: how *well* does our model predict the response?

Class Activity, Part I

Predictions with Titanic data:

https://sta214-s23.github.io/class_activities/ca_lecture_14.html

$\hat{Y}_i =$ predicted outcome (0 or 1)

Class activity

Fitted model:

$$\log\left(\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}\right) = 3.78 - 0.037 Age_i - 2.52 Male_i - 1.31 Class2_i - 2.58 Class3_i$$

> What is the predicted probability of survival for a male, first-class passenger aged 20? What about for a male, second-class passenger aged 30?

① $$\hat{\pi}_i = \frac{\exp(3.78 - 20(0.037) - 2.52)}{1 + \exp(3.78 - 20(0.037) - 2.52)} \approx 0.63$$

$\hat{Y}_i = 1$

② $$\hat{\pi}_i = \frac{\exp(3.78 - 30(0.037) - 2.52 - 1.31)}{1 + \exp(3.78 - 20(0.037) - 2.52 - 1.31)} \approx 0.24$$

$\hat{Y}_i = 0$

Making predictions with the Titanic data

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

> How do we turn $\widehat{\pi}_i$ into a binary prediction of survival / no survival?

$\hat{\pi}_i$ = predicted probability

$\hat{Y}_i$ = predicted outcome

$$\hat{Y}_i = \begin{cases} 1 & \hat{\pi}_i \geq 0.5 \\ 0 & \hat{\pi}_i < 0.5 \end{cases}$$

threshold

Confusion matrix

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

table(Prediction = m1$fitted.values > 0.5,
      Truth = titanic$Survived)
```

$\hat{\pi}_i$ threshold Y

```
##            Truth
## Prediction   0   1
##      FALSE 356  83
##      TRUE   68 207
```

$Y=0$ $Y=1$

$\hat{Y}=0$ (FALSE), $\hat{Y}=1$ (TRUE)

356: # passengers where $Y=0$ and $\hat{Y}=0$

83: $\hat{Y}=0$ but $Y=1$

68: $\hat{Y}=1$ but $Y=0$

207: $\hat{Y}=1$ and $Y=1$

> Did we do a good job predicting survival?

Ideally, want accuracy, sensitivity, specificity, PPV all high

Confusion matrix

$\hat{Y}=0$ $\hat{Y}=1$

```
##             Truth
## Prediction    0    1
##      FALSE  356   83
##      TRUE    68  207
```

	$Y=0$	$Y=1$
$\hat{Y}=0$	True negative (TN)	False negative (FN)
$\hat{Y}=1$	False positive (FP)	True positive (TP)

Accuracy : $\dfrac{\#\text{ correct predictions}}{\#\text{ observations}} = \dfrac{TN + TP}{n} = \dfrac{356+207}{714} \approx 0.79$

Sensitivity : $P(\hat{Y}=1 \mid Y=1) = \dfrac{TP}{TP+FN} = \dfrac{207}{207+83} = 0.71$

Specificity : $P(\hat{Y}=0 \mid Y=0) = \dfrac{TN}{TN+FP} = \dfrac{356}{356+68} = 0.84$

Positive predictive value (PPV) : $P(Y=1 \mid \hat{Y}=1) = \dfrac{TP}{TP+FP} = \dfrac{207}{207+68} = 0.75$

Negative predictive value : $P(Y=0 \mid \hat{Y}=0)$

Class activity, Part II

Predictions with the SBA data:

https://sta214-s23.github.io/class_activities/ca_lecture_14.html

Class activity

```
m1 <- glm(Default ~ log(DisbursementGross) + Term +
            as.factor(UrbanRural),
          family = binomial, data = sba)

table(Prediction = m1$fitted.values > 0.5,
      Truth = sba$Default)
```

```
##            Truth
## Prediction FALSE TRUE
##      FALSE  3989  734
##      TRUE    100  168
```

- Accuracy = (3989 + 168) / 4991 = 0.83
- Sensitivity = 168/(168 + 734) = 0.19
- Specificity = 3989/(3989 + 100) = 0.98
- PPV = 168/(168+100) = 0.63

Class activity

```
##             Truth
## Prediction FALSE TRUE
##      FALSE  3989  734
##      TRUE    100  168
```

> Is an accuracy of around 80% good?

It depends on proportion of 0s and 1s in the data

E.g., consider:

	$Y=0$	$Y=1$
$\hat{Y}=0$	4089	902
$\hat{Y}=1$	0	0

Accuracy ≈ 0.82

By itself, accuracy is meaningless if we have imbalanced data

. Accuracy is highest when threshold ≈ 0.5

Class activity

. As threshold $\uparrow$, sensitivity $\downarrow$, specifity $\uparrow$

Changing thresholds: trade-off between sensitivity & specificity

```
table(Prediction = m1$fitted.values > 0.3,
      Truth = sba$Default)
```

```
##             Truth
## Prediction FALSE TRUE
##      FALSE  3524  351
##      TRUE    565  551
```

Accuracy = 0.82
Sensitivity = 0.61
Specificity = 0.86

```
table(Prediction = m1$fitted.values > 0.7,
      Truth = sba$Default)
```

```
##             Truth
## Prediction FALSE TRUE
##    ŷ=0 FALSE  4089  902
```

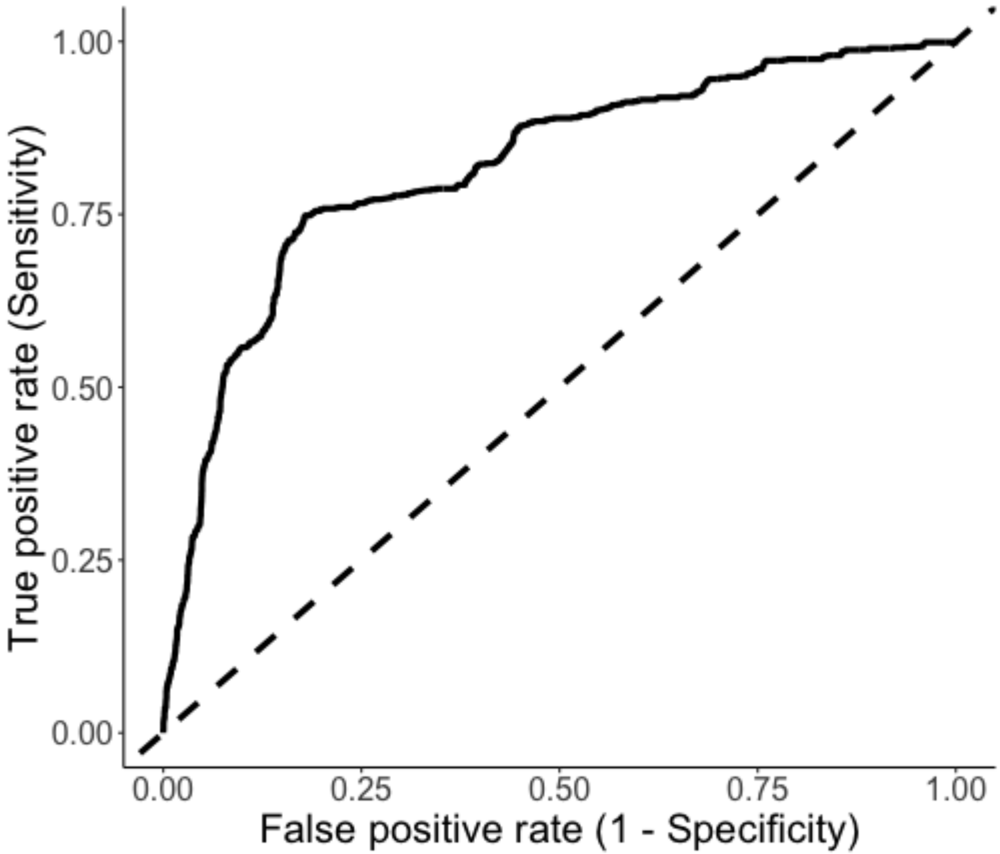
$\hat{y}=1$ 0 0

Accuracy = 0.82
Sensitivity = 0
Specificity = 1

Changing thresholds

> How can I assess prediction performance across many different thresholds?

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)