106-1 生物統計學二 實習課

R : ROC curve

周芷妤

d04849010@ntu.edu.tw

2017.11.30

大綱

- ROC curve
  - Fit logistic model & calculate $\hat{p}_X$
  - Scatter plot & classification boundary
  - Calculate sensitivities & specificities
  - Plot ROC curve
    - Calculate AUC
    - Find optimal cut point
  - Calculate misclassification rate
  - Histogram

ROC curve

Fit logistic model & calculate $\hat{p}_X$
Scatter plot & classification boundary
Calculate sensitivities & specificities
Plot ROC curve
Calculate misclassification rate
Histogram

Logistic model & $\hat{p}_X$

- Fit logistic model & calculate $\hat{p}_X = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2}}$

✓ model <- glm(Y ~ X1 + X2 , family = binomial , data = 資料檔名稱)

✓ 利用predict指令: predict(model, type="response")

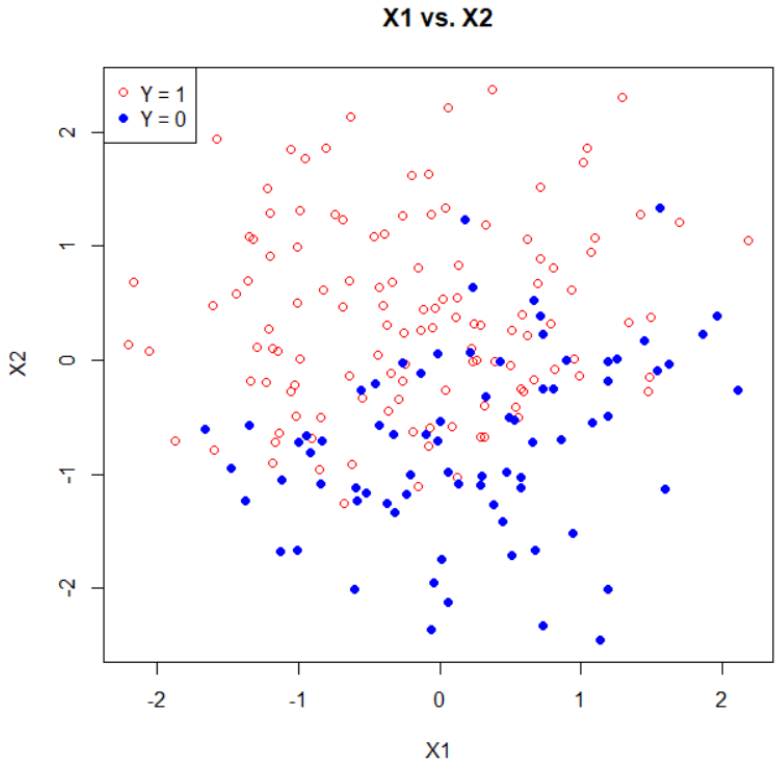
```
> model.1 <- glm(Y~ X1+X2, family = binomial, data = data_hw6 )
> pred.value <- predict(model.1, type="response")
```

也可以利用公式計算$\hat{p}_X$: exp(…)/(1+exp(…))

Scatter plot

- Plot the data points (X1,X2) with different symbols for Y = 0 and Y = 1.

```
> plot(X1, X2, pch = 1, main="X1 vs. X2")
> points(X1[Y==1],X2[Y==1], col="red", pch=1)
> points(X1[Y==0],X2[Y==0], col="blue", pch=16)
> legend("topleft", c("Y = 1", "Y = 0"), col=c("red", "blue"), pch=c(1, 16))
```

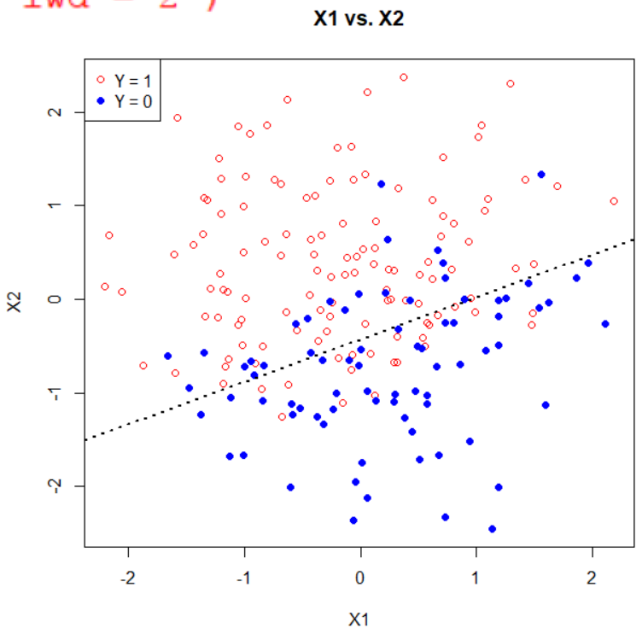


Classification boundary

- Add classification boundary on scatter plot with prediction rule $\{\hat{p}_X > 0.5\}$

$$\ln\left\{\frac{\hat{p}_X}{1-\hat{p}_X}\right\} = \ln\left\{\frac{0.5}{1-0.5}\right\} = 0 = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \quad \Longrightarrow \quad X_2 = -\frac{\hat{\beta}_0}{\hat{\beta}_2} - \frac{\hat{\beta}_1}{\hat{\beta}_2} X_1$$

```
> b.coef <- model.1$coefficients
> intercept.cb <- - b.coef[1]/b.coef[3]
> slope.cb <- - b.coef[2]/b.coef[3]
>
> abline(intercept.cb, slope.cb, lty = 3, lwd = 2 )
```



Sensitivities & Specificities

- Calculate sensitivity & specificity
  - Prediction rule $\{\hat{p}_X > c\}$:

分子: $Y=1$的那一類中，$\hat{p}_X$大於c的個數
分母: $Y=1$的個數

$$se(c) = \mathrm{P}(\hat{p}_X > c | Y=1) = \frac{\text{number of } (\hat{p}_X > c \text{ and } Y=1)}{\text{number of } (Y=1)}$$

$$sp(c) = \mathrm{P}(\hat{p}_X < c | Y=0) = \frac{\text{number of } (\hat{p}_X < c \text{ and } Y=0)}{\text{number of } (Y=0)}$$

→ 給定一個c，就得到一個$se(c)$和一個$sp(c)$

```
> se.1 <- c()
> sp.1 <- c()
> c.seq <- seq(0, 1, 0.0001)
>
> for(ii in 1:length(c.seq)){
+
+ se.1[ii] <-  sum( pred.value[Y==1] > c.seq[ii] )  / sum(Y==1)
+ sp.1[ii] <-  sum( pred.value[Y==0] < c.seq[ii] )  / sum(Y==0)
+
+ }
```

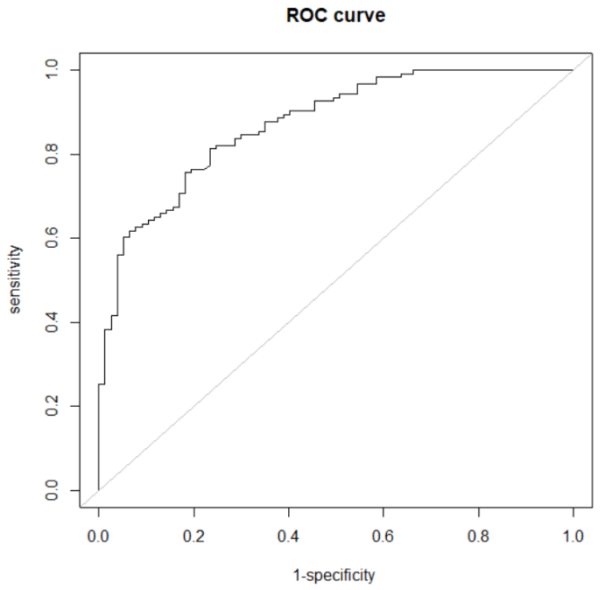
定義物件空間

ROC curve

- Plot ROC curve

```
> plot(1-sp.1, se.1, xlim = c(0,1), ylim = c(0,1), type="l",
+      main = "ROC curve", xlab = "1-specificity ", ylab = "sensitivity" )
>
> abline(0,1, col = "grey")
```

加上最沒有預測能力的ROC curve (45度線)
✓ abline(截距, 斜率)

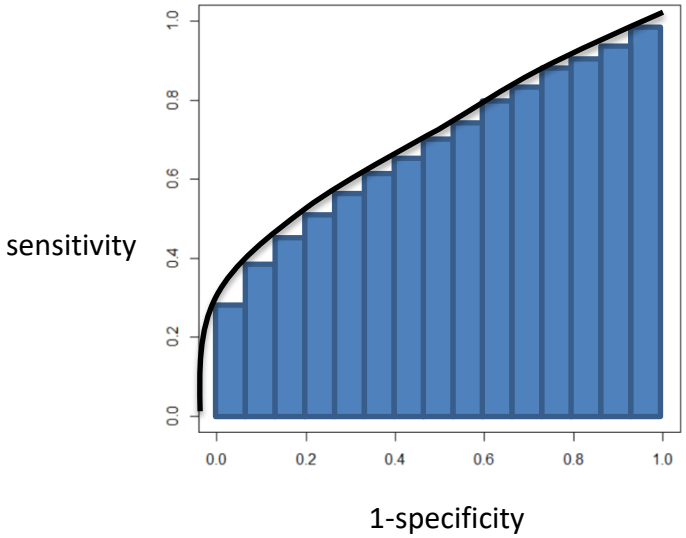


AUC

- Calculate AUC

→ 計算曲線下面積

概念:計算每個長方形的面積後加總




```
> auc.value <- c()
> for(ii in 1:length(c.seq)-1){
+
+ height.auc <- se.1[ii+1]
+ width.auc <- abs((1-sp.1[ii])- (1-sp.1[ii+1]))
+
+ auc.value[ii] <- height.auc*width.auc
+
+ }
>
> sum(auc.value)
```

計算每個長方形的面積

將每個長方形的面積加總

Optimal cut point

- Find the optimal cut point c^*

  Youden index : $\max\{se(c) + sp(c)\}$

optimal ←

c	specificity	sensitivity
c_1	$sp(c_1)$	$se(c_1)$
⋮	⋮	⋮
$\boldsymbol{c^*}$	$\boldsymbol{sp(c^*)}$	$\boldsymbol{se(c^*)}$
⋮	⋮	⋮
c_k	$sp(c_k)$	$se(c_k)$

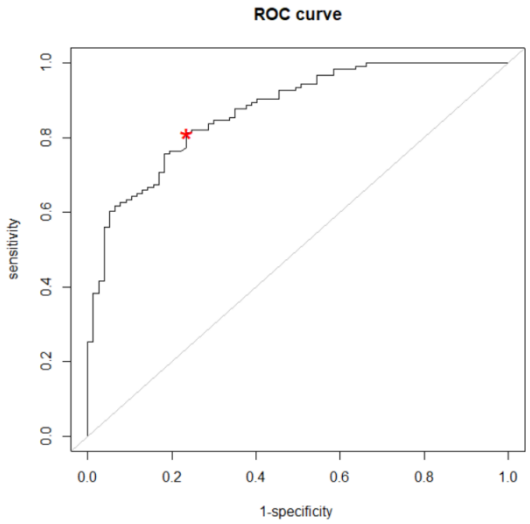
```
> youden.index <- max(se.1 + sp.1)
> optimal.cutpoint <- which.max(se.1 + sp.1)
> c.seq[optimal.cutpoint]
```

找出c^*的位置

找出c^*

Add point & text

- Add ($1 - sp(c^*)$,$se(c^*)$) on ROC curve

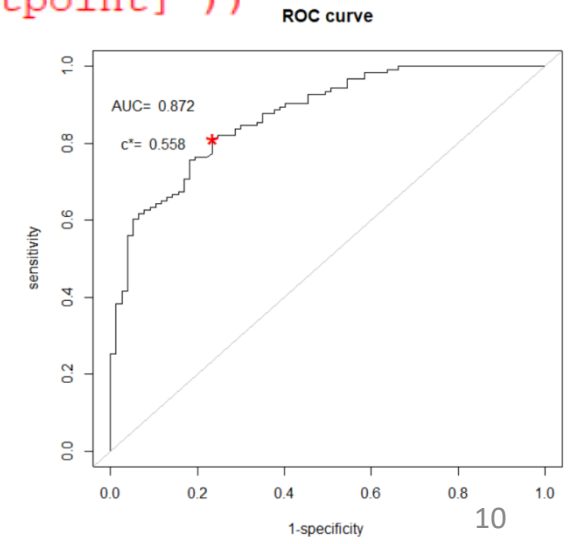



```
> se.opt <- se.1[optimal.cutpoint]
> sp.opt <- sp.1[optimal.cutpoint]
>
> points(1-sp.opt, se.opt, pch = "*", cex = 2.5, col = "red")
```

- Add text on ROC curve

```
> text(0.1, 0.9, paste("AUC= ", round(sum(auc.value),3) ))
> text(0.1, 0.8, paste("c*= ", c.seq[optimal.cutpoint] ))
```

在圖上加AUC和c^*的數值
- ✓ paste(…): 串聯字串和計算數值
- ✓ round(…, 3): 計算到小數第三位



Misclassification rate

- Calculate misclassification rate

$$\mathrm{MR}_{(c)} = \frac{\text{number of } (Y_i \neq \hat{Y}_i)}{n}$$

```
> se.05 <-  sum( pred.value[Y==1] > 0.5 )  / sum(Y==1)
> sp.05 <-  sum( pred.value[Y==0] < 0.5 )  / sum(Y==0)
> mis.rate.05 <-  ( (1-se.05)*sum(Y==1) + (1-sp.05)*sum(Y==0) ) / length(Y)
>
>
> mis.rate.youden <-  ( (1-se.opt)*sum(Y==1) + (1-sp.opt)*sum(Y==0) ) / length(Y)
```

計算$\mathrm{MR}_{(0.5)}$

計算$\mathrm{MR}_{(c^*)}$

Histogram

- Prediction rule $\{ f(X_1) > c \}$ with $f(X_1) = \frac{e^{X_1}}{1+e^{X_1}}$:

利用plogis計算$\frac{e^{X_1}}{1+e^{X_1}}$

在長方形中以密度為15，畫45度的斜線

```
> pred.value8 <- plogis( X1 )

> hist(pred.value8[Y==1], xlim=c(0,1), main="histogram of f(X1)",
+      xlab="f(X1)", ylab="frequency", col="red", density= 15, angle= 45)
> hist(pred.value8[Y==0], add=T, col="blue", density= 15, angle= 135)
> abline(v=0.5, col="black", lwd=2)
> legend( "topright", c("Y = 1", "Y = 0"), col=c("red", "blue"), lty=1 )
```