

Lab 5

Biostatistics 210

0. BACKGROUND

This lab is meant to give you practice in comparing models.

1. DOWNLOAD

•Download the model building dataset `lab-learn.dta`. The data was collected by the United Network on Organ Sharing and contains 1 year survival outcomes on pediatric kidney transplants carried out in the US between 1990 and 2002. The variables are

- death (1=died in first year, 0 = survived first year) — the outcome
- age (age of child at transplant)
- age_don (age of the kidney donor)
- txtype (1=cadaveric donor, 0 = living donor)
- prevtx (1= previous transplant, 0=otherwise)
- year (calendar year of transplant)

2. FITTING THE MODELS

The is the same dataset you used in lab 4. In that lab, AIC and BIC suggested a model with predictors: year txtype age. The cross validated AUC suggested year txtype age and prevtx. In this lab, we will look more closely at their fit

First fit the 4 predictor model

```
logistic death year txtype age prevtx
```

and save the predicted probabilities in “out4”

```
predict out4, p
```

Then fit the 3 predictor model

```
logistic death year txtype age
```

and save the predicted probabilities in “out3”

It is helpful to see how they differ. Just try a scatter plot of the predictions from the two models

```
twoway (scatter out3 out4) xtitle(3 Variable Model) ytitle(4
Variable Model)
```

3. DISCRIMINATION GRAPH

Now see if you can detect a difference in their discrimination by looking at if the predicted probabilities are very different between those who live and those who died.

```
twoway (kdensity out4 if death==1, bw(.01)) (kdensity out3 if
death==1, bw(.01)) (kdensity out4 if death==0, bw(.01)) (kden-
sity out3 if death==0, bw(.01)), legend(order( 1 "4 Variable -
Died" 2 "3 Variable - Died" 3 "4 Variable - Lived" 4 "3 Variable
- Died"))
```

to do the above in Stata — you will need to tweak the quote marks

Do you see much a difference in the predictions among those who live — among those who die?

4. NET RECLASSIFICATION

Suppose, we classify children with a greater than 0.075 (or 7.5%) predicted chance of death as high risk. This will prioritize specificity over sensitivity and ensure a high positive predictive value of the high risk classification.

```
gen highrisk4 = (out4 > 0.075)
gen highrisk3 = (out3 > 0.075)
label def hr 1 "High Risk" 0 "Not High Risk"
label val highrisk4 highrisk3 hr
```

Calculate the classification of high risk children among those who died?

```
tabulate highrisk4 highrisk3 if death==1
```

What is the sensitivity of this classification for "highrisk4"?

What is the sensitivity of this classification for "highrisk3"?

What is the net reclassification?

Calculate the classification of high risk children among those who live?

tabulate highrisk4 highrisk3 if death==0

What is the sensitivity of this classification for "highrisk4"?

What is the sensitivity of this classification for "highrisk3"?

What is the net reclassification?

Which model would you prefer?