

Lab 4

Biostatistics 210

0. BACKGROUND

This lab is meant to give you practice in predictor selection in the context of trying to build a model which is likely to give good prediction in the future.

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)

•Also, download the do file `lab-cv.do`

2. PREDICTOR SELECTION

The dataset has 5 predictors (age, age_don, txtype, prevtx, year) and it is easy to explore a large number of possible models. Note that the total number of possible models is $2^5 = 32$ which is not feasible to explore. Instead, do a series of backward selection models. Trying models with 5 predictors, 4 predictors, etc. At each step, delete the variable which has the largest p-value (equivalently the z-statistic which has a value which is farthest from zero). Don't categorize the continuous variables for this lab and don't worry about the "*xi: i.*" syntax

As you fit each model, record the AIC/BIC. For a given model, you can obtain the AIC and BIC by typing

```
logistic death yoursetofpredictors
```

```
estat ic
```

Your first model should be: `logistic death age age_don year txtype prevtx.`

To obtain the area under ROC curve you will need to open the do file editor for lab-cv.-do. This series of commands creates 10 mutually exclusive subsets of data. It fits the model of leaving one out (at a time) and then predicts that group. It does this until each of the 10 groups is left out.

The do file allows you to calculate this directly but you will need to modify one section of the file. The part in red below should be replaced by the list of predictors for the model you are testing.

```
forvalues k=1(1)10 {
  quietly xi: logistic death <yoursetofpredictors> if cat!=`k'
  quietly predict out`k'
  quietly replace pred=out`k' if cat==`k'
}
```

Complete the table below.

# of Variables	AIC	BIC	Area under ROC (CV)
5			
4			
3			
2			
1			

Discuss which model appears optimal based on the minimizing the BIC and (CV) area under the ROC curve. Which model would you choose as likely to yield good predictions for a new set of observations?