

# Lab #7

## *Biostatistics 210*

### 0. BACKGROUND

This lab is meant to give you practice in evaluating and validating a prognostic model. We will use an adapted version of last weeks dataset. .

### 1. DOWNLOAD

Download the model building dataset lab7.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. This weeks data consists of two parts: a training set which you used last week and a validation set. The variables are

- death (1=died in first year, 0 = survived first year)
- age (age of child at transplant)
- agecode (1= age 0 to 3, 0 = age 4 to 18)
- txtype (1=cadaveric donor, 0 = living donor)
- year (calendar year of transplant)
- yrcode (1 = year 1990-1998, 0 = year 1999-2002)
- valid (1=validation set, 0 = training set)

## 2. PREDICTION MODEL

Use as your prediction model the variables: tctype, agecode, and yrccode. The categorization of the continuous predictions helps to make a very simple coding which takes integer values. The decision of how to code these was made to ensure that the points for each of the variables is positive.

Recall, you can get the points for each of the variables by calculating the coefficients for each variable (on the training set) and dividing by the smallest coefficient

```
logistic death tctype agecode yrccode if valid==0, coef
```

Save the linear predictor

```
predict linear_pred, xb
```

and predicted probability

```
predict prob, p
```

How many points do you calculate for

1. tctype
2. agecode
3. yrccode

Generate a variable called "points" using the generate command based on the number of points you have given to each variable.

## 3. MODEL DISCRIMINATION

You can calculate the area under the ROC associated with the scoring with the commands

Test set:

```
quietly logistic death points if valid==0  
lroc
```

Validation set

```
quietly logistic death points if valid==1  
lroc
```

In addition you might examine the boxplots of the points variable among those who lived and died in the learning and validation set.

```
graph box points, over(death) over(valid) scheme(s1 color) ytitle("Score")
```

**Q1: Given the ROC and boxplots, what is the predictive power of the model?**

#### 4. MODEL CALIBRATION

Calculate the observed and expected mortality for each score in the validation set and in the test set

```
table points if valid==1,c(n prob mean prob mean d)format(%9.2f)
```

```
table points if valid==0,c(n prob mean prob mean d)format(%9.2f)
```

Examine the lowess curve of the predicted probability versus mortality in the test and validation set

```
twoway (lowess death prob if valid==1) (lowess death prob if valid==0) (lfit prob prob),  
legend(order(1 "Validation" 2 "Learning" )) scheme(s1 color) ytitle("Observed Probability")  
xtitle("Estimated Probability")
```

**Q2: How well calibrated does the model appear in the test set and the validation set**