Practical: RISK PREDICTION
Advanced Statistics for Records Research

Research question

In this session, we will explore the dataset of 2000 participants we met in the lecture, and fit a risk prediction model for death within 5 years, based on some simple patient characteristics.

Objectives

By the end of this practical, you should be able to:

1. Fit a logistic model to create risk predictions.

2. Assess model discrimination by calculating the Area Under the Curve.

3. Assess model calibration by graphing observed and predicted risks.

Dataset and analysis

For this practical we will use a (simulated) dataset called "**Prediction-data.dta**". This contains data for 2,000 patients, with information on six variables.

Variable	Description
id	Unique patient ID
age	Age (years)
sbp	Systolic Blood Pressure
bmi	Body Max Index kg/m2
sex	Female=0, Male=1
dead	Alive=0, Dead=1

Preparatory steps

1. Copy the dataset "**Prediction-data.dta**" to your own folder.

2. Open a do-file. Start by changing the working directory using the `cd` command

3. Open the dataset "**Prediction-data.dta**".

Data exploration

4. Have a look at the data. How many participants die? What proportion are female?

What ages are these participants?

Randomly split data into training and validation parts

5. Create a variable S = 0/1 that separates the data into two equal halves:

```
. set seed 1111
. gen rvar = runiform()
. sort rvar
. gen S = 0
. replace S =1 if _n>1000
```

Fit model in training data and predict risks

6. For the S=0 dataset, estimate the model using logistic regression. Include all measured variables as predictors.

7. Predict the risk of death for all individuals (i.e. those with S=0 and those with S=1).

```
. predict m2pr
```

Validation

8. Draw an ROC curve in the S=0 dataset.

```
. roctab dead m2pr if S==0, graph specificity
```

What is the AUC? Interpret this number. Now repeat for the S=1 dataset. Are the two ROC curves very different?

9. Create a Hosmer-Lemeshow goodness of fit table for the S=0 and the S=1 datasets.

```
. estat gof if S==0, group(10) table
```

10. Draw a bar graph comparing the predicted and observed risks in the S=0 data.

```
. egen m2prg0 = cut(m2pr) if S==0, group(10)
. graph bar (sum) m2pr (sum) dead, over(m2prg0)
```

Repeat for the S=1 dataset.

11. If you have time, see if you can write a short do-file to repeatedly split the data, calculate the AUC from the S=0 and S=1 datasets and store the results.