Practical session 7: RISK PREDICTION

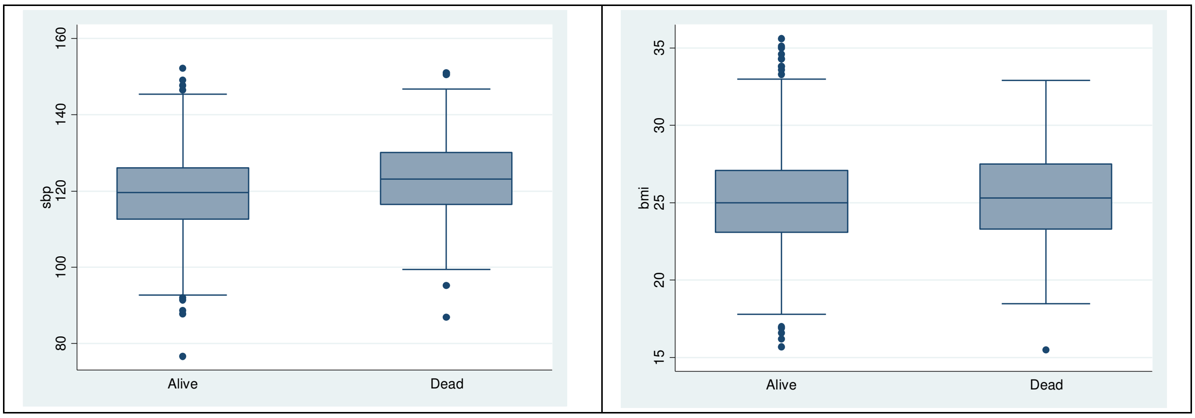
Advanced Statistics for Records Research

SOLUTIONS

Data exploration

4. About 25% of the 2,000 individuals die – this is a high-risk population. There is roughly 50% males and 50% females, aged 40-80.

```
. tab dead

      event |      Freq.     Percent        Cum.
------------+-----------------------------------
      Alive |      1,491       74.55       74.55
       Dead |        509       25.45      100.00
------------+-----------------------------------
      Total |      2,000      100.00

. tab sex

        sex |      Freq.     Percent        Cum.
------------+-----------------------------------
     Female |        978       48.90       48.90
       Male |      1,022       51.10      100.00
------------+-----------------------------------
      Total |      2,000      100.00

. summ age

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |      2,000      60.451    11.54423         40         80
```



Randomly split data into training and validation parts

5. [No output, except a newly created indicator variable S]

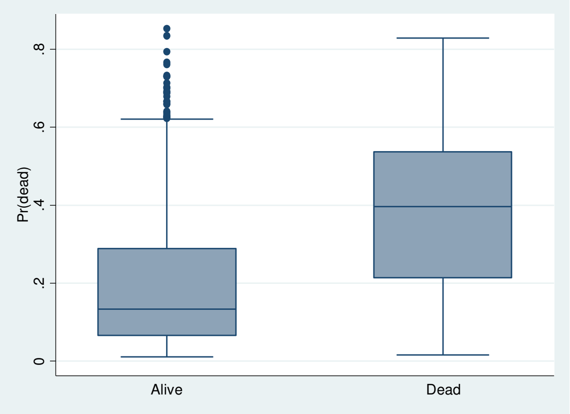
Fit model in training data and predict risks

6. The results below are for the variable S created after setting the seed to 1111 (in Stata version 14.2).

```
. logistic dead c.age i.sex c.sbp c.bmi  if S==0

Logistic regression                             Number of obs     =      1,000
                                                LR chi2(4)        =     205.95
                                                Prob > chi2       =     0.0000
Log likelihood = -442.27278                     Pseudo R2         =     0.1889

------------------------------------------------------------------------------
        dead | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    1.10318   .0093403    11.60   0.000     1.085025     1.12164
             |
         sex |
       Male  |   1.299215   .2178909     1.56   0.119     .9352468    1.804828
         sbp |   1.046004   .0091674     5.13   0.000      1.02819    1.064127
         bmi |   1.051433   .0289298     1.82   0.068     .9962331    1.109691
       _cons |   6.62e-07   9.74e-07    -9.67   0.000     3.70e-08    .0000118
------------------------------------------------------------------------------
```

7. The predicted risks are summarised and graphed below (in datasets S=0 and S=1 combined).

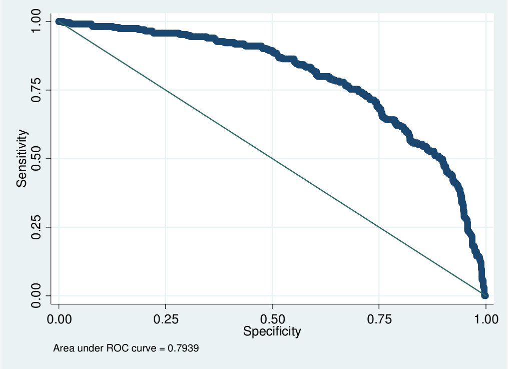
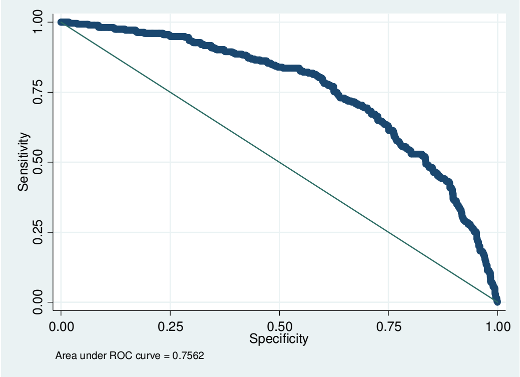
```
. bysort dead: summ m2pr
-------------------------------------------------------------------------------
-> dead = Alive

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        m2pr |      1,491     .193726    .1632586   .0114158   .8532454

-------------------------------------------------------------------------------
-> dead = Dead

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        m2pr |        509    .3846043    .1976729   .0166571    .828801

. graph box m2pr, over(dead)
```



Validation

8. In the training dataset, the ROC is 79%. This means that a person who did die has a 79% probability of having a higher predicted risk (of dying) than someone who did not. This shows the model has fairly good discrimination (ability to separate those who did and did not experience the event of interest).

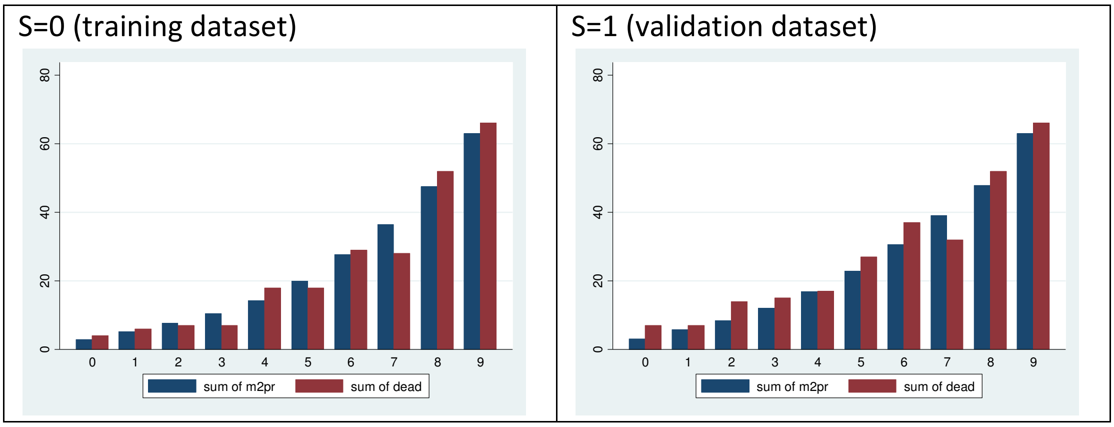
```
. roctab dead m2pr if S==0, graph specificity
```

S=0 (training dataset)	S=1 (validation dataset)
	

9. The Hosmer-Lemeshow goodness of fit table for the two (S=0 and the S=1) datasets were very similar. Both showed evidence of a well calibrated model.

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

Logistic model for dead, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.0410 |     4 |   2.9 |    96 |  97.1 |   100 |
  |     2 | 0.0630 |     6 |   5.1 |    94 |  94.9 |   100 |
  |     3 | 0.0909 |     7 |   7.6 |    93 |  92.4 |   100 |
  |     4 | 0.1206 |     7 |  10.5 |    93 |  89.5 |   100 |
  |     5 | 0.1678 |    18 |  14.3 |    82 |  85.7 |   100 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     6 | 0.2295 |    18 |  19.9 |    82 |  80.1 |   100 |
  |     7 | 0.3214 |    29 |  27.6 |    71 |  72.4 |   100 |
  |     8 | 0.4164 |    28 |  36.4 |    72 |  63.6 |   100 |
  |     9 | 0.5331 |    52 |  47.5 |    48 |  52.5 |   100 |
  |    10 | 0.8532 |    66 |  63.1 |    34 |  36.9 |   100 |
  +--------------------------------------------------------+

       number of observations =      1000
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =         7.66
                  Prob > chi2 =         0.4671
```

10. Bar graphs comparing the predicted and observed risks in the S=0 and S=1 datasets also show good calibration (in both the training and validation data).



S=0 (training dataset)

S=1 (validation dataset)