

SOLUTIONS

Practical: RISK PREDICTION Advanced Statistics for Records Research

2. Change working directory
3. Open the dataset "**Prediction-data.dta**"

[click 'File > Open' then choose "Prediction-data.dta"]

Data exploration

4. Have a look at the data.

summarize

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-------|----------|-----------|------|-------|
| id | 2,000 | 1000.5 | 577.4946 | 1 | 2000 |
| age | 2,000 | 60.451 | 11.54423 | 40 | 80 |
| sbp | 2,000 | 120.3289 | 10.18013 | 76.6 | 152.2 |
| bmi | 2,000 | 25.188 | 2.982409 | 15.5 | 35.6 |
| sex | 2,000 | .511 | .500004 | 0 | 1 |
| dead | 2,000 | .2545 | .4356887 | 0 | 1 |

How many participants die?

tab dead

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

What proportion are female?

tab sex

| sex | Freq. | Percent | Cum. |
|--------|-------|---------|-------|
| Female | 978 | 48.90 | 48.90 |

| | | | | |
|-------------|--|-------|--------|--------|
| Male | | 1,022 | 51.10 | 100.00 |
| -----+----- | | | | |
| Total | | 2,000 | 100.00 | |

What ages are these participants?

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. 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

(1,000 real changes made)

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

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.

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

```
Logistic regression                                Number of obs   =      1,000
                                                    LR chi2(4)      =      205.31
                                                    Prob > chi2     =      0.0000
Log likelihood = -452.99024                        Pseudo R2       =      0.1848
```

| dead | Odds Ratio | Std. Err. | z | P> z | [95% Conf. Interval] | |
|-------|------------|-----------|--------|-------|----------------------|----------|
| age | 1.093615 | .0088335 | 11.08 | 0.000 | 1.076438 | 1.111066 |
| sex | | | | | | |
| Male | 1.076061 | .177804 | 0.44 | 0.657 | .7783727 | 1.487599 |
| sbp | 1.066765 | .0094303 | 7.31 | 0.000 | 1.048441 | 1.085409 |
| bmi | 1.044692 | .0286565 | 1.59 | 0.111 | .9900095 | 1.102395 |
| _cons | 1.43e-07 | 2.13e-07 | -10.60 | 0.000 | 7.76e-09 | 2.64e-06 |

Note: _cons estimates baseline odds.

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

predict m2pr

bysort dead: summ m2pr

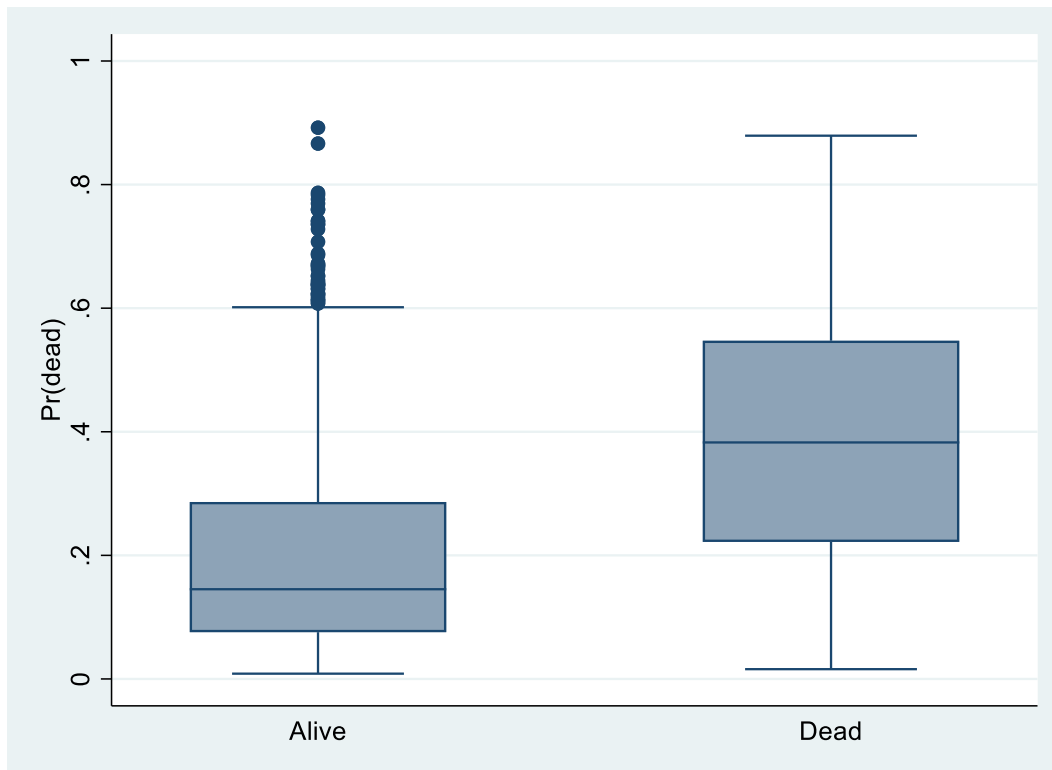
-> dead = Alive

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-------|----------|-----------|----------|----------|
| m2pr | 1,491 | .1984497 | .1615049 | .0085352 | .8920841 |

-> dead = Dead

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-----|----------|-----------|----------|----------|
| m2pr | 509 | .3901502 | .2043112 | .0158026 | .8791203 |

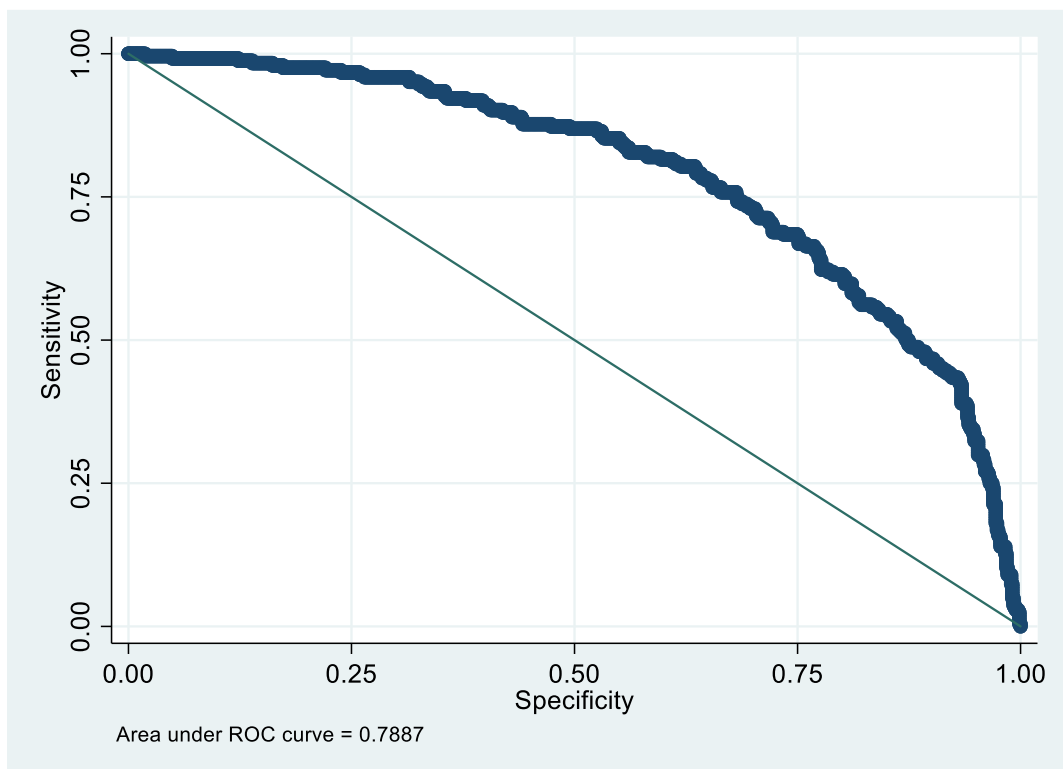
graph box m2pr, over(dead)



Validation

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

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

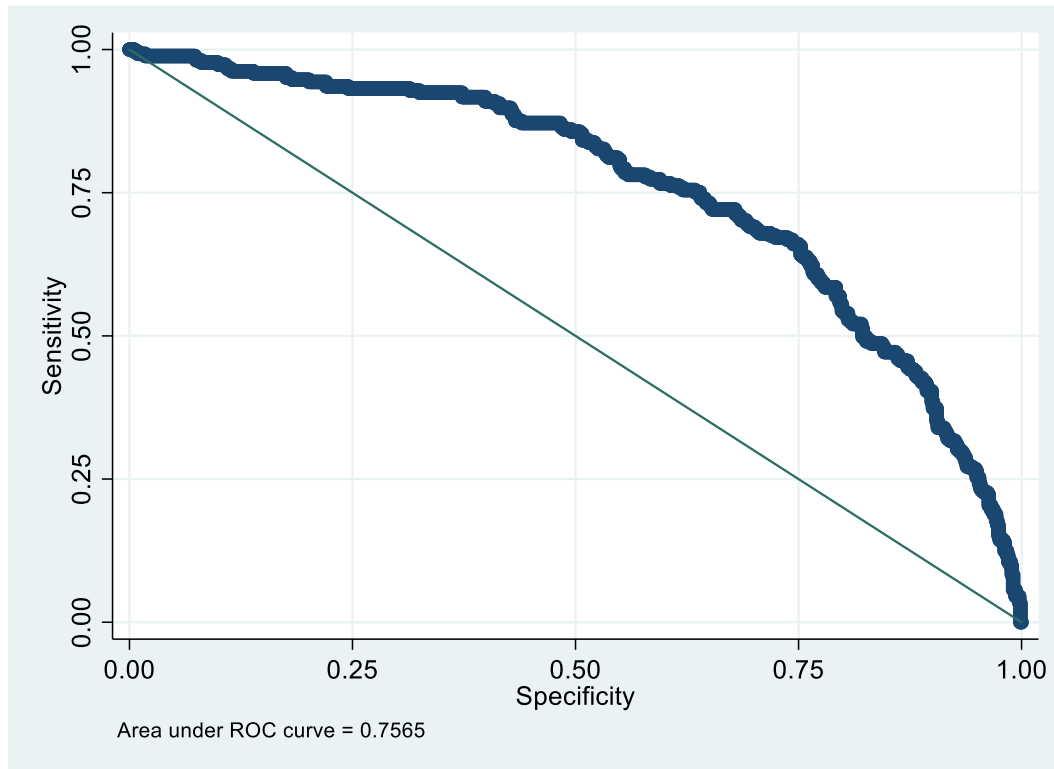


What is the AUC? Interpret this number.

In the training set, 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).

Now repeat for the S=1 dataset.

roctab dead m2pr if S==1, graph specificity



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

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 | 0.0470 | 3 | 3.1 | 97 | 96.9 | 100 |
| 2 | 0.0725 | 5 | 6.0 | 95 | 94.0 | 100 |
| 3 | 0.0992 | 11 | 8.6 | 89 | 91.4 | 100 |
| 4 | 0.1327 | 12 | 11.4 | 88 | 88.6 | 100 |
| 5 | 0.1813 | 14 | 15.5 | 86 | 84.5 | 100 |
| 6 | 0.2422 | 23 | 21.3 | 77 | 78.7 | 100 |
| 7 | 0.3230 | 27 | 28.0 | 73 | 72.0 | 100 |
| 8 | 0.4272 | 32 | 36.9 | 68 | 63.1 | 100 |
| 9 | 0.5374 | 48 | 47.8 | 52 | 52.2 | 100 |
| 10 | 0.8791 | 69 | 65.2 | 31 | 34.8 | 100 |

number of observations = 1000
 number of groups = 10
 Hosmer-Lemeshow chi2(8) = 2.99
 Prob > chi2 = 0.9348

estat gof if S==1, 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 | 0.0449 | 10 | 3.0 | 90 | 97.0 | 100 |
| 2 | 0.0777 | 8 | 6.2 | 92 | 93.8 | 100 |
| 3 | 0.1081 | 4 | 9.3 | 96 | 90.7 | 100 |
| 4 | 0.1507 | 15 | 13.0 | 85 | 87.0 | 100 |
| 5 | 0.1960 | 25 | 17.1 | 75 | 82.9 | 100 |
| 6 | 0.2596 | 21 | 22.7 | 79 | 77.3 | 100 |
| 7 | 0.3316 | 32 | 29.9 | 68 | 70.1 | 100 |
| 8 | 0.4249 | 36 | 37.5 | 64 | 62.5 | 100 |
| 9 | 0.5390 | 49 | 47.4 | 51 | 52.6 | 100 |
| 10 | 0.8921 | 65 | 64.3 | 35 | 35.7 | 100 |

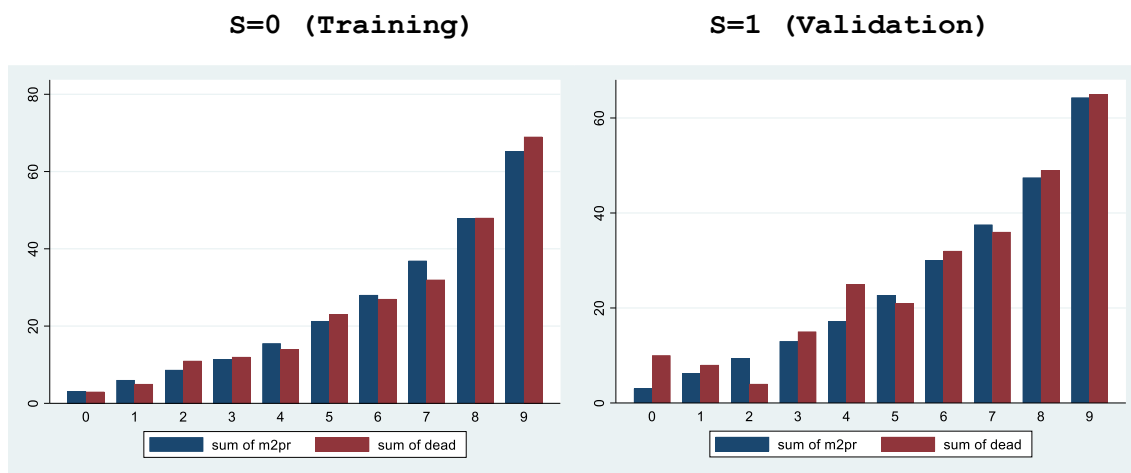
number of observations = 1000
 number of groups = 10
 Hosmer-Lemeshow chi2(8) = 26.12
 Prob > chi2 = 0.0010

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.

```
egen m2prg1 = cut(m2pr) if S==1, group(10)
graph bar (sum) m2pr (sum) dead, over(m2prg1)
```



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.