Risk Prediction

Advanced Statistics for Records Research

Luigi Palla

Based on slides by David Prieto and Elizabeth Williamson

Medical Statistics Department, London School of Hygiene & Tropical Medicine, Farr Institute of Health Informatics, London

Learning objectives

By the end of this lecture, you will be able to:

- Describe the difference between classification and risk prediction
- Explain how ROC curves are calculated and how they relate to sensitivity and specificity
- Assess goodness-of-fit by graphing observed and predicted risks

The problem

Consider individuals in which a certain binary event Y might happen (0=no, 1=yes) by a certain point in time. We also have a set of other variables for each individual. Some examples:

- A group of patients about to have a surgery: Some will die in the surgery and some will survive. We have data on their age, sex, severity of disease, comorbidities etc.
- A group of pregnant women: some will have a child with a malformation, some will not. We have data on the mothers' age, medications, diagnostics, life-style, etc.
- A group of patients who got a hip replacement, some will fail and some won't, we have data on demographics, biomarkers and lifestyle.

In all these cases we do not know for sure which individuals will get the event.

What we want to do

Ideally we would like to predict who is going to get the event. This can be done in two ways:

- **Deterministic:**
  You classify each individual in one of two categories: either you think they will have the event or you think they will not. (This deterministic **Classification** is used in *Machine Learning*).

- **Probabilistic:**
  You don't just classify people as either yes or no, you assign each person a probability of having the event. (This is often used in *biostatistics* and is known as **Risk Prediction**).

For either of these two strategies you use the data that you have collected from the individual (age, sex, diagnoses, etc. . .).

MI in the next 5 years?

- Suppose these are our data and we want to predict who will have a myocardial infarction in the next 5 years

ID	Age	Sex	Diabet	SBP	Classify	Predict	Observed
1	35	F	1	145			
2	35	M	0	130			
3	55	F	0	115			
4	55	M	1	170			
5	65	F	0	135			
6	65	M	1	140			
7	75	M	1	160			
8	75	F	0	130			
9	85	F	1	130			
10	85	M	0	160			

Classify Deterministically

- Use a previous algorithm: C(age, sex, diabet, SBP) = 0 or 1
- C(35, F, 1, 145) = 0, . . . , C(55, M, 1, 170) = 1

ID	Age	Sex	Diabet	SBP	Classify	Predict	Observed
1	35	F	1	145	**No**		
2	35	M	0	130	**No**		
3	55	F	0	115	**No**		
4	55	M	1	170	**Yes**		
5	65	F	0	135	**No**		
6	65	M	1	140	**Yes**		
7	75	M	1	160	**Yes**		
8	75	F	0	130	**No**		
9	85	F	1	130	**Yes**		
10	85	M	0	160	**Yes**		

Predict Probabilistically

- Use a previous algorithm: P(age, sex, diabet, SBP) = [0, 1]
- P(35, F, 1, 145) = 0.15, . . . , P(55, M, 1, 170) = 0.55.

ID	Age	Sex	Diabet	SBP	Classify	Predict	Observed
1	35	F	1	145		**0.15**	
2	35	M	0	130		**0.05**	
3	55	F	0	115		**0.10**	
4	55	M	1	170		**0.55**	
5	65	F	0	135		**0.30**	
6	65	M	1	140		**0.52**	
7	75	M	1	160		**0.60**	
8	75	F	0	130		**0.40**	
9	85	F	1	130		**0.55**	
10	85	M	0	160		**0.60**	

Observe

- … After 5 years follow-up
- Now you can compare predictions with observed events

ID	Age	Sex	Diabet	SBP	Classify	Predict	Observed
1	35	F	1	145	No	0.15	**No**
2	35	M	0	130	No	0.05	**No**
3	55	F	0	115	No	0.10	**No**
4	55	M	1	170	Yes	0.55	**No**
5	65	F	0	135	No	0.30	**Yes**
6	65	M	1	140	Yes	0.52	**No**
7	75	M	1	160	Yes	0.60	**Yes**
8	75	F	0	130	No	0.40	**No**
9	85	F	1	130	Yes	0.55	**Yes**
10	85	M	0	160	Yes	0.60	**Yes**

Validating our algorithms/models

- Our models have predicted some individuals well but not others.
- We want to have a general measure of how good or bad our algorithms are by comparing the predictions with the actual observed values.
- This will be done differently for a classification algorithm than for a prediction algorithm.

Sensitivity and specificity

Sensitivity

- Probability of correctly predicting the cases (MI).
- Given you are someone who will have an MI, what is the probability of a positive classification?

Specificity

- Probability of correctly predicting the non-cases (no MI)
- Given you are someone who will not have an MI, what's the probability of a negative classification?

Truth

Classification

Data	Negative	Positive	Total
Negative	A	B	(A+B)
Positive	C	D	(C+D)
Total	(A+C)	(B+D)	(A+B+C+D)

Sensitivity: D/(B+D)

Specificity: A/(A+C)

Properties of the test

Predictive value

Positive predictive value (PPV)

- Probability of disease given a positive classification

Negative predictive value (NPV)

- Probability of no disease given a negative classification

Affected by prevalence of disease

Truth

Classification

Data	Negative	Positive	Total
Negative	A	B	(A+B)
Positive	C	D	(C+D)
Total	(A+C)	(B+D)	(A+B+C+D)

PPV: D/(C+D)

NPV: A/(A+B)

Comparing classifications with observations

Truth (observed)

Classification

Data	No	Yes	Total
No	4	1	5
Yes	2	3	5
Total	6	4	10

Sensitivity:

- 3/4 = 75%

Specificity:

- 4/6 = 66%

Positive predictive value:

- 3/5 = 60%

Negative predictive value:

- 4/5 = 80%

Comparing risk predictions with observations

- Order individuals by risk and choose a cut-off point to classify them as “YES” only if the prediction exceeds the cut-off. For example:
  - ➢ **Classify as “Yes”" if Prediction >0.1**

ID	Predict	Observed	Yes if Pred>0.1
2	0.05	No	**No**
3	0.10	No	**No**
1	0.15	No	**Yes**
5	0.30	Yes	**Yes**
8	0.40	No	**Yes**
6	0.52	No	**Yes**
4	0.55	No	**Yes**
9	0.55	Yes	**Yes**
7	0.60	Yes	**Yes**
10	0.60	Yes	**Yes**

Cut-off point: “Yes” if P>0.1

Truth (observed)

Classification

Data	No	Yes	Total
No	2	0	2
Yes	4	4	8
Total	6	4	10

Sensitivity:

- 4/4 = 100%

Specificity:

- 2/6 = 33%

Positive predictive value:

- 4/8 = 50%

Negative predictive value:

- 2/2 = 100%

- Higher sensitivity but less specificity than the ML classification algorithm.

Cut-off point: “Yes” if P>0.4

- Ty a different cut-off point:
  - **Classify as “Yes”" if Prediction > 0.4**

ID	Predict	Observed	Yes if Pred>0.4
2	0.05	No	**No**
3	0.10	No	**No**
1	0.15	No	**No**
5	0.30	Yes	**No**
8	0.40	No	**No**
6	0.52	No	**Yes**
4	0.55	No	**Yes**
9	0.55	Yes	**Yes**
7	0.60	Yes	**Yes**
10	0.60	Yes	**Yes**

Cut-off point: “Yes” if P>0.4

Truth (observed)

Classification

Data	No	Yes	Total
No	4	1	5
Yes	2	3	5
Total	6	4	10

Sensitivity:

- ➢ 3/4 = 75%

Specificity:

- ➢ 4/6 = 66%

Positive predictive value:

- ➢ 3/5 = 60%

Negative predictive value:

- ➢ 4/5 = 80%

- Same comparison table as for the ML classification algorithm.

Cut-off point: “Yes” if P>0.55

- Ty a different cut-off point:
  - **Classify as “Yes”" if Prediction > 0.55**

ID	Predict	Observed	Yes if Pred>0.55
2	0.05	No	**No**
3	0.10	No	**No**
1	0.15	No	**No**
5	0.30	Yes	**No**
8	0.40	No	**No**
6	0.52	No	**No**
4	0.55	No	**No**
9	0.55	Yes	**No**
7	0.60	Yes	**Yes**
10	0.60	Yes	**Yes**

Cut-off point: “Yes” if P>0.55

Truth (observed)

Classification

Data	No	Yes	Total
No	6	2	8
Yes	0	2	2
Total	6	4	10

Sensitivity:

- 2/4 = 50%

Specificity:

- 6/6 = 100%

Positive predictive value:

- 2/2 = 100%

Negative predictive value:

- 6/8 = 75%

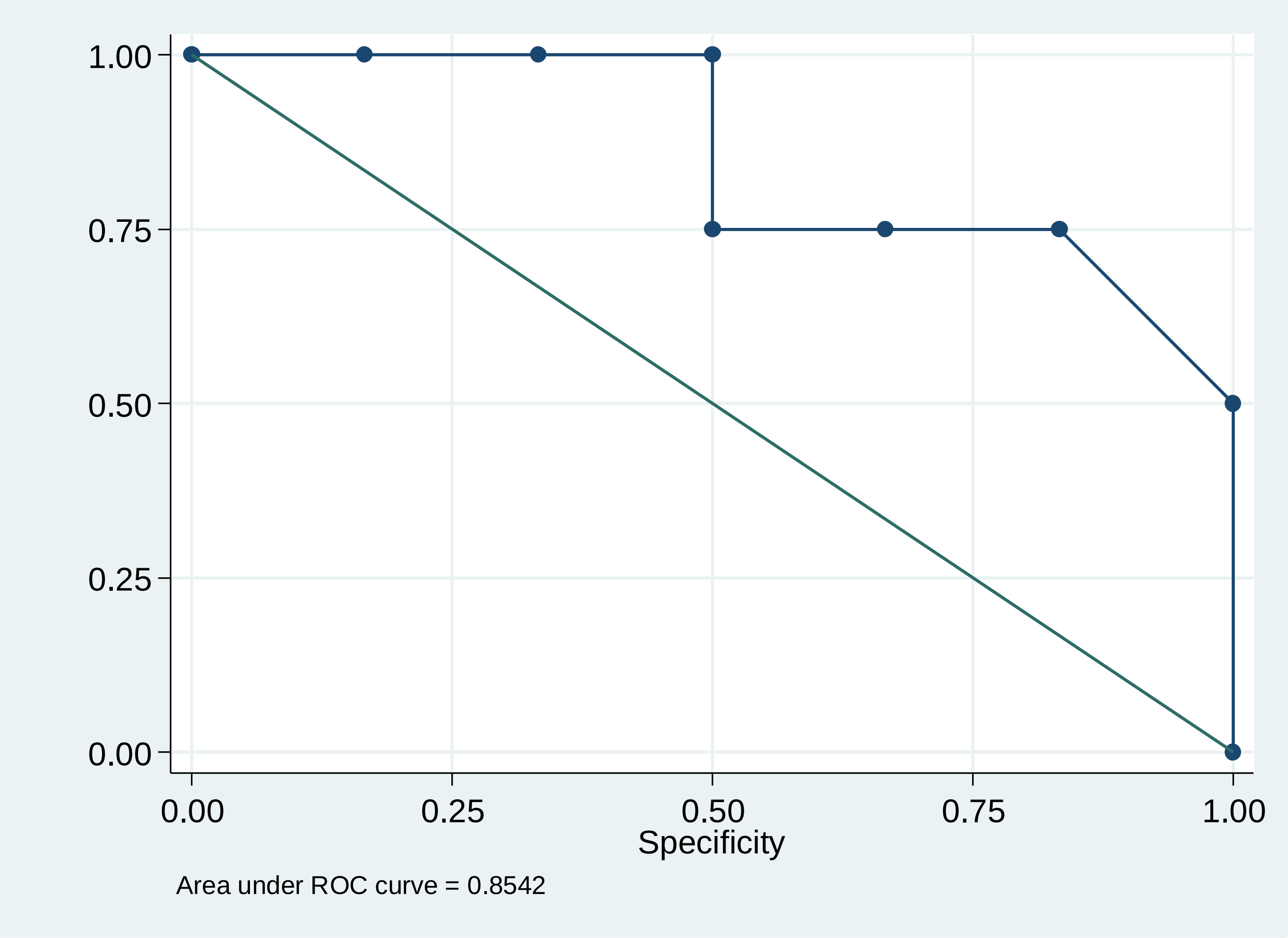
- Lower sensitivity, higher specificity than the ML classification algorithm.

All cut-offs

- If we repeat this process for each change in the predictive value in the table we will obtain a list of sensitivities and specificities.

ID	Predict	Observed	Cut-off	Sensit.	Specif.
2	0.05	No	P>0.05	100	17
3	0.10	No	P>0.10	100	33
1	0.15	No	P>0.15	100	50
5	0.30	Yes	P>0.30	75	50
8	0.40	No	P>0.40	75	66
6	0.52	No	P>0.52	75	83
4	0.55	No			
9	0.55	Yes	P>0.55	50	100
7	0.60	Yes			
10	0.60	Yes	P>0.60	0	100

Receiver Operator Characteristic (ROC) Curve



A curve linking all the sensitivities against the specificities (in the table above).

The Area Under the Curve (AUC)

- AUC can be interpreted as the probability that an observed “yes” was assigned a higher probability than an observed “no”.
- If the prediction model was useless (same as assigning probabilities “at random”), then an observed “yes” would have only 50% chances of having higher predictive risk than an observed “no” (AUC = 0.5)
- AUC would be 1 for a method that would give higher predictions to all the observed “yes” than to all the observed “no” (perfect separation between “yes” and “no”).
- Any real-world model will have their AUC between 0.5 and 1. Closer to 1 indicates better performance in separating cases and controls.

How do we come up with the predictions?

- We propose a statistical model for the probability of the event happening $P(Y_i = 1)$ depending on the other variables and some coefficients.
- For example a logistic model:

$$\log\left(\frac{P(Y_i=1)}{1-P(Y_i=1)}\right) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \cdots \qquad (1)$$

- We need a "training set" where we can observe all the variables Y_i, X_i, Z_i, ... to estimate the coefficients β_0, β_1, β_2,...
- Once we have the coefficients that best fit the data we can calculate the predicted risk for each individual "i"

$$\hat{P}(Y_i = 1) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_i+\hat{\beta}_2 Z_i+\cdots}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_i+\hat{\beta}_2 Z_i+\cdots}} \qquad (2)$$

Validating the Prediction Model

Internal validation:

- We compare the predicted risks in each individual with the actual observed events with classification tables and ROC curves.

(Semi-) external validation:

- We use the coefficients β_0, β_1, β_2,…, to predict the risk in a DIFFERENT set of individuals not used for the estimation. If those predictions seem to coincide with their own events then the model seems to be valid externally.
- If we only have one database we can divide it randomly in two sets: one to estimate the model and the other to validate it.
- To be even more sure you can repeat this many times in your dataset by doing a different random partition each time and re-estimating the model and the validation exercise. The final model will be a sort of average of the models.

A larger example with 2000 individuals

- We will use the variables Age, Sex, SBP, and BMI to predict if the person will be Dead=0 or Alive=1 in 5 years time.

```
| id     Age   SBP      BMI     Sex       Death |
|-----------------------------------------------|
|   1    47    116.8    25.6    Female    Alive |
|   2    71    113.8    21.1    Male      Dead  |
|   3    41    130.7    25.6    Male      Alive |
|   4    71    118.2    25.9    Male      Alive |
|   5    54    120.3    20.7    Male      Alive |
|   6    67    126.4    22.8    Male      Alive |
|   7    71    129.3    32.9    Female    Dead  |
|   8    73    117.4    27.5    Male      Alive |
|   ...  ..    .....    .....   ......    ..... |
|   2000 46    111.1    24      Male      Alive |
+-----------------------------------------------+
```

Logistic Regression; Model M1

- We could fit a model to all the data to estimate the coefficients.
- Note that BMI is not very significant (P=0.1885), but let’s stay with this model for now

```
. logistic dead age sex sbp bmi

Logistic regression                                   Number of obs = 2,000
LR chi2(4)  = 377.64
Prob > chi2 = 0.0000
Log likelihood = -945.62964                              Pseudo R2 = 0.1664
------------------------------------------------------------------------------
        dead |  Odds Ratio  Std. Err.      z     P>|z|    [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  1.09391    .0062246    15.77   0.000   1.081778   1.106178
         sex |  1.301071   .1484571     2.31   0.021   1.040341   1.627145
         sbp |  1.050522   .0061981     8.35   0.000   1.038444   1.062741
         bmi |  1.025531   .0195045     1.33   0.185   .9880064   1.06448
       _cons|  1.39e-06    1.38e-06  -13.52   0.000   1.96e-07   9.81e-06
------------------------------------------------------------------------------
```

Create Predictions from Model M1

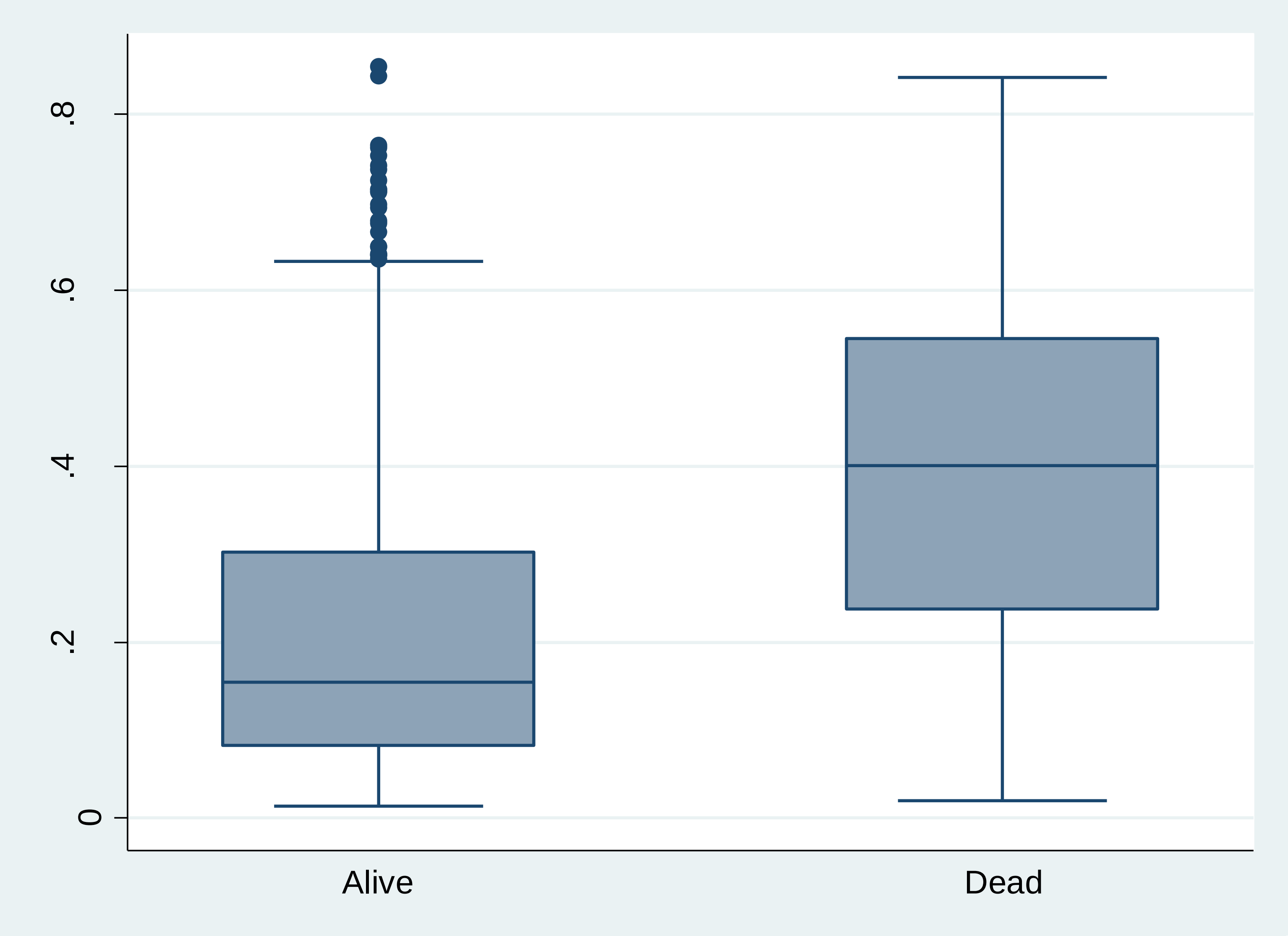
- This command creates a variable with the linear predictor i.e. the logit of the probability of event as in equation (1)

```
. predict m1lp, xb
```

- This command creates a variable with the predicted probability of death as in equation (2)

```
. predict m1pr
+------------------------------------------------------------------+
| id   age   sbp    bmi   sex     dead    m1lp          m1pr       |
|------------------------------------------------------------------|
| 10   46   111.1   24    Male    Alive  -3.014752     .0467639    |
| 11   79   117.5   28.7  Male    Dead     .381197     .594162     |
| 12   72   113.2   24.2  Male    Alive   -.572492     .3606619    |
| 13   54   123.4   26.1  Female  Alive  -1.900696     .1300298    |
| 14   45   105.7   24.7  Female  Alive  -3.616203     .0261807    |
```

Predictions in dead and alive



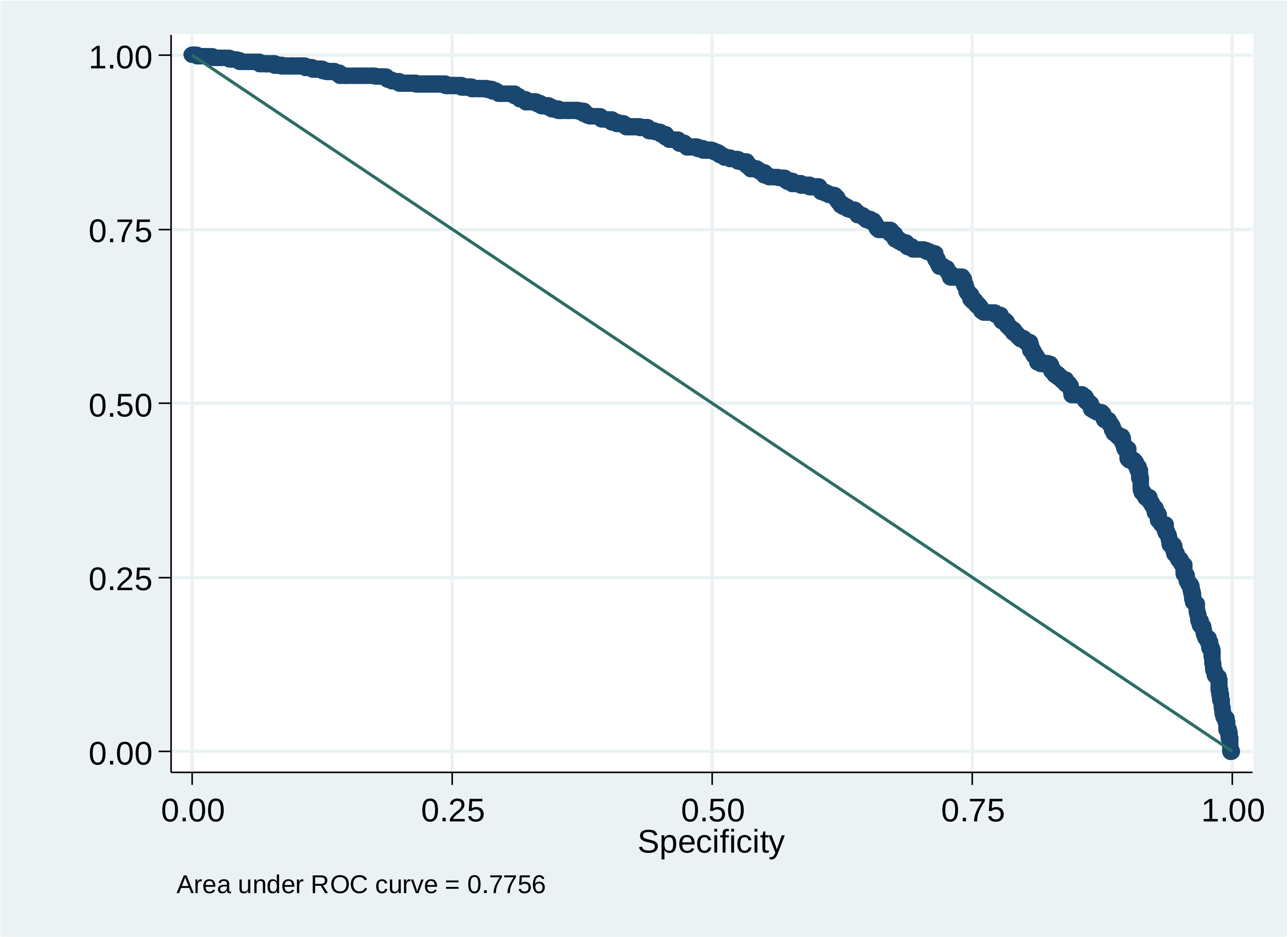
Classification table: cutoff P(Y=1) >= 0.3

```
. estat classification, cutoff(0.3)
---------------------- True ----------
  Classified |      D          ~D      |    Total
-------------+-------------------------+----------
      +      |     333        375      |     708
      -      |     176       1116      |    1292
-------------+-------------------------+----------
    Total    |     509       1491      |    2000

Classified + if predicted Pr(D) >= .3
----------------------------------------------
Sensitivity                     Pr( +| D)   65.42%
Specificity                     Pr( -|~D)   74.85%
Positive predictive value       Pr( D| +)   47.03%
Negative predictive value       Pr(~D| -)   86.38%
False + rate for classified +   Pr(~D| +)   52.97%
False - rate for classified -   Pr( D| -)   13.62%
Correctly classified                        72.45%
----------------------------------------------
```

Classification table: cutoff P(Y=1) >= 0.5

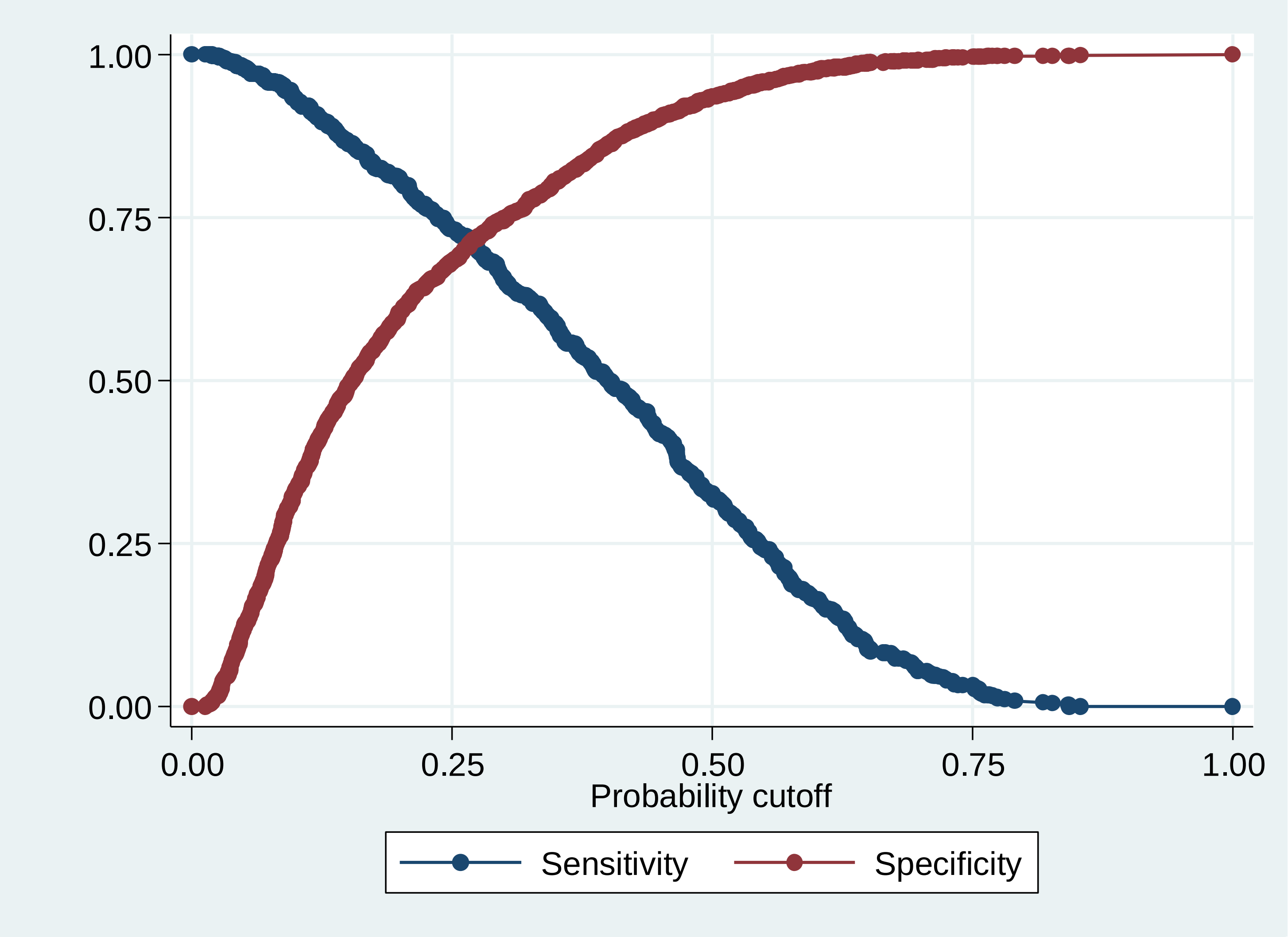
```
. estat classification, cutoff(0.5)
------------------------ True ------------
   Classified |       D         ~D       |    Total
--------------+--------------------------+-----------
       +      |      166        96       |     262
       -      |      343       1395      |    1738
--------------+--------------------------+-----------
     Total    |      509       1491      |    2000

Classified + if predicted Pr(D) >= .5
--------------------------------------------------
Sensitivity                     Pr( +| D)  32.61%
Specificity                     Pr( -|~D)  93.56%
Positive predictive value       Pr( D| +)  63.36%
Negative predictive value       Pr(~D| -)  80.26%
False + rate for classified +   Pr(~D| +)  36.64%
False - rate for classified -   Pr( D| -)  19.74%
Correctly classified                       78.05%
--------------------------------------------------
```

ROC curve from model M1



- The Area Under the Curve (AUC) is 0.78
- There is a 78% probability that a person that actually dies gets a higher predicted risk by the model than a person that did not die by the end of the follow up.

Sensitivity and Specificity by cut-off value for model M1



- In practice, need to choose a cut-off for clinical decision-making
- Can plot Sensitivity-Specificity against cut-off value.
- This can help to select the most convenient cut-off depending on whether we need more sensitivity or specificity in our problem.

For discussion

1) High-sensitivity, low-specificity

2) Low-sensitivity, high-specificity

- Ebola Disease Virus
- Prostate cancer in elderly men
- Meningococcal meningitis

For discussion

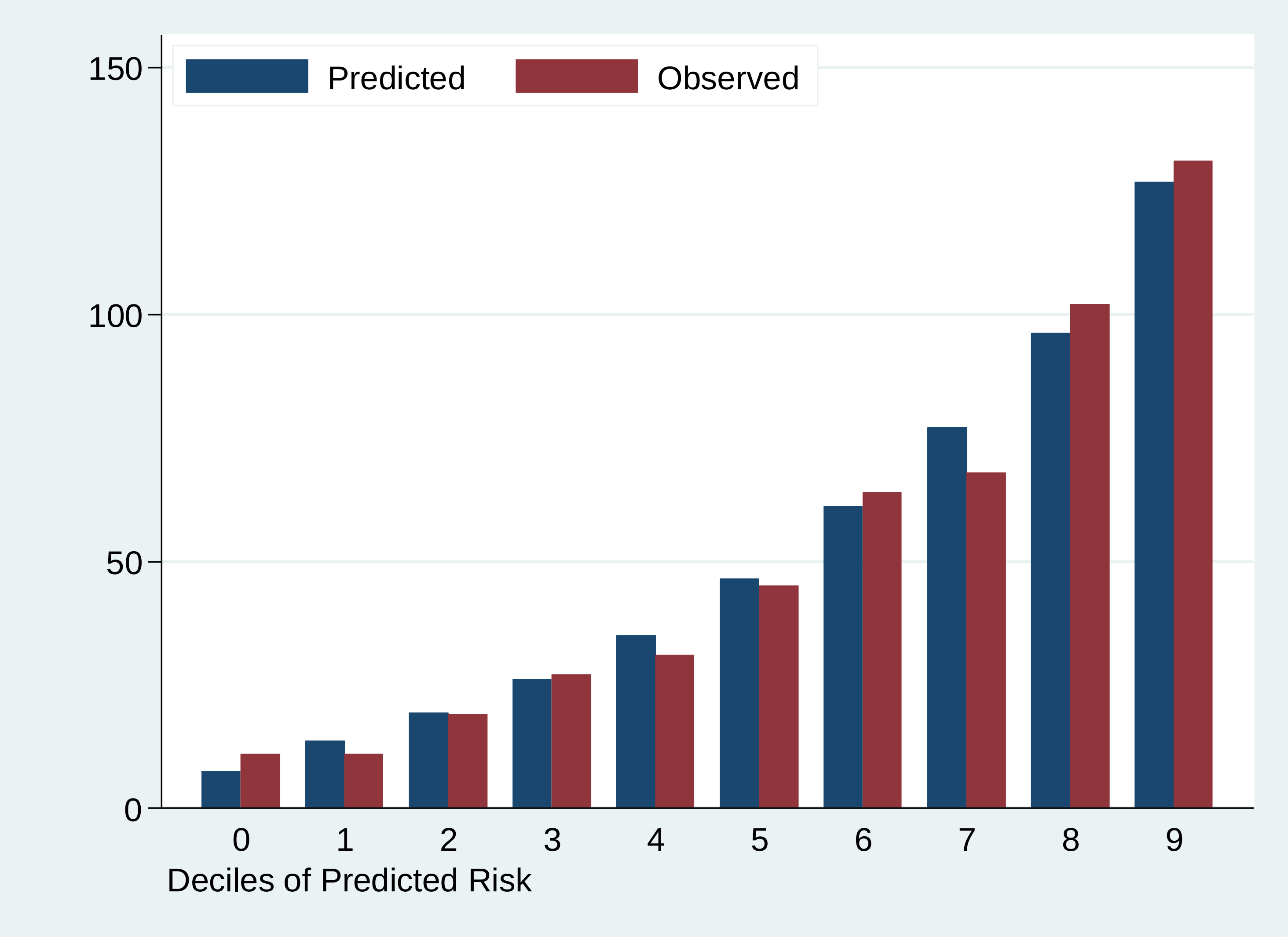
1) High-sensitivity, low-specificity	a) More false-positives (+ve test, no disease)
2) Low-sensitivity, high-specificity	b) More false-negatives (-ve test, disease)

- Ebola Disease Virus
- Prostate cancer in elderly men
- Meningococcal meningitis

Goodness of fit: Hosmer-Lemeshow test

```
. estat gof, 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   | 0.0520 |   11   |   7.4  |  189  | 192.6 |  200  |
|   2   | 0.0831 |   11   |  13.6  |  189  | 186.4 |  200  |
|   3   | 0.1133 |   19   |  19.3  |  181  | 180.7 |  200  |
|   4   | 0.1504 |   27   |  26.1  |  173  | 173.9 |  200  |
|   5   | 0.2011 |   31   |  35.0  |  169  | 165.0 |  200  |
|   6   | 0.2670 |   45   |  46.4  |  155  | 153.6 |  200  |
|   7   | 0.3460 |   64   |  61.1  |  136  | 138.9 |  200  |
|   8   | 0.4277 |   68   |  77.0  |  132  | 123.0 |  200  |
|   9   | 0.5384 |  102   |  96.2  |   98  | 103.8 |  200  |
|   10  | 0.8544 |  131   | 126.7  |   69  |  73.3 |  200  |
+------------------------------------------------------+
Hosmer-Lemeshow chi2(8) = 5.98 Prob > chi2 = 0.6499
```

Goodness-of-fit: Observed and Expected events by deciles of risk



Model validation

Key measures in validation of the model…

- **Discrimination**
  - The ability of the model to distinguish between patients who have the event (MI) and don’t (no MI)
  - Often assessed by AUC
- **Calibration (goodness-of-fit)**
  - The agreement between the observed & predicted outcomes
  - For a group of patients with 10% predicted risk, do 10% experience the event?
  - E.g. graph in previous slide
- **Clinical usefulness**
  - Does the model provide accurate predictions at the patient level that can be used to guide clinical decision making?
  - E.g. Decision analysis

Summary

- Risk prediction models (binary outcomes) often estimated via logistic regression
- AUC is a useful measure of the model discrimination
- Comparing observed and predicted risks is a useful way to assess calibration
- Other approaches, e.g. decision analysis, are required to assess clinical benefit to individual patients

Connections with ML approaches

- Breiman L. Statistical Modeling: The Two Cultures. Statistical Science, 2001, 16 (3):199–231 [with discussion]

- Key points:
  - Two cultures in the use of statistical modelling:
    1. Assume data are generated by a given stochastic data model [today]
    2. Use algorithmic models and treat data mechanism as unknown [machine learning]
  - "If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools"
    - E.g. neural networks, random forests, support vector machines..
- Very interesting discussion

References and further reading

- Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. European Heart Journal, 2014, 35:1925–1931. doi:10.1093/eurheartj/ehu207

- Breiman L. Statistical Modeling: The Two Cultures. Statistical Science, 2001, 16 (3):199–231 [with discussion]