

## STAT 504

## Analysis of Discrete Data

## 7.2.3 - Receiver Operating Characteristic Curve (ROC)

 [Printer-friendly version \(https://onlinecourses.science.psu.edu/stat504/print/book/export/html/163\)](https://onlinecourses.science.psu.edu/stat504/print/book/export/html/163)

A Receiver Operating Characteristic Curve (ROC) is a standard technique for summarizing classifier performance over a range of trade-offs between true positive (TP) and false positive (FP) error rates (Sweets, 1988). ROC curve is a plot of *sensitivity* (*the ability of the model to predict an event correctly*) versus *1-specificity* for the possible cut-off classification probability values  $\pi_0$ .

For logistic regression you can create a  $2 \times 2$  *classification table* of predicted values from your model for your response if  $\hat{y} = 0$  or 1 versus the true value of  $y = 0$  or 1. The prediction if  $\hat{y} = 1$  depends on some cut-off probability,  $\pi_0$ . For example,  $\hat{y} = 1$  if  $\hat{\pi}_i > \pi_0$  and  $\hat{y} = 0$  if  $\hat{\pi}_i \leq \pi_0$ . The most common value for  $\pi_0 = 0.5$ .

Then *sensitivity* =  $P(\hat{y} = 1|y = 1)$  and *specificity* =  $P(\hat{y} = 0|y = 0)$ .

The ROC curve is more informative than the classification table since it summarizes the predictive power for all possible  $\pi_0$ .

The position of the ROC on the graph reflects the accuracy of the diagnostic test. It covers all possible thresholds (cut-off points). The ROC of random guessing lies on the diagonal line. The ROC of a perfect diagnostic technique is a point at the upper left corner of the graph, where the TP proportion is 1.0 and the FP proportion is 0.

The Area Under the Curve (AUC), also referred to as index of accuracy (A), or *concordance index*,  $c$ , in SAS, and it is an accepted traditional performance metric for a ROC curve. The higher the area under the curve the better prediction power the model has.  $c = 0.8$  can be interpreted to mean that a randomly selected individual from the positive group has a test value larger than that for a randomly chosen individual from the negative group 80 percent of the time.

For more details see Agresti(2007), Sections 5.1.6-5.1.8, Agresti (2013), Section 6.2, and SAS links provided at the beginning.

Using SAS (#)

Using R (#)



Here is the SAS program assay4.sas

(/stat504/sites/onlinecourses.science.psu.edu.stat504/files/lesson06/assay4.sas) .

```
options nocenter nodate nonumber linesize=72;
data assay;
  input logconc y n;
  cards;
2.68 10 31
2.76 17 30
2.82 12 31
2.90 7 27
3.02 23 26
3.04 22 30
3.13 29 31
3.20 29 30
3.21 23 30
;
run;

proc logistic data=assay;
  model y/n= logconc / scale=pearson outroc=roc1;
  output out=out1 xbeta=xb reschi=reschi;
run;

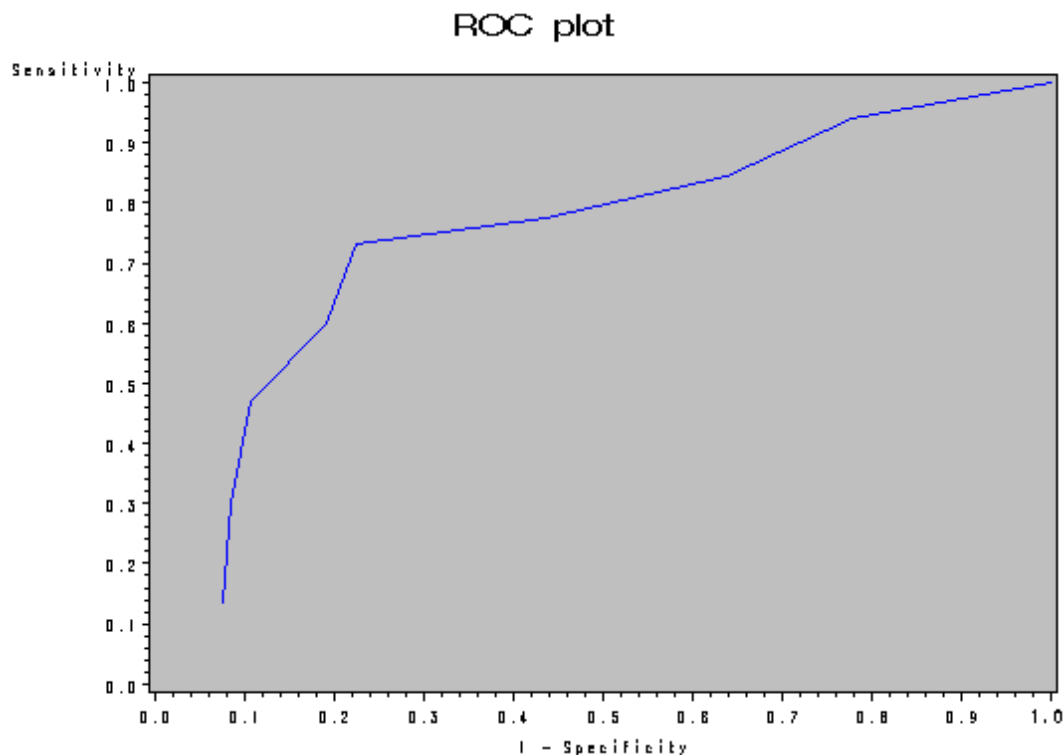
axis1 label=('Linear predictor');
axis2 label=('Pearson Residual');

proc gplot data=out1;
  title 'Residual plot';
  plot reschi * xb / haxis=axis1 vaxis=axis2;
run;

symbol1 i=join v=none c=blue;

proc gplot data=roc1;
  title 'ROC plot';
  plot _sensit_ *_lmspec_=1 / vaxis=0 to 1 by .1 cframe=ligr ;
run;
```

Here is the resulting ROC graph.



Area under the curve is  $c = 0.746$  indicates good predictive power of the model.

#### Association of Predicted Probabilities and Observed Responses

Percent Concordant	70.6	Somers' D	0.492
Percent Discordant	21.4	Gamma	0.535
Percent Tied	8.0	Tau-a	0.226
Pairs	16168	c	0.746

Option **ctable** prints the classification tables for various cut-off points. Each row of this output is a classification table for the specified **Prob Level**,  $\pi_0$ .

Classification Table

Prob Level	Correct		Incorrect		Percentages				
	Event	Non- Event	Event	Non- Event	Correct	Sensi- tivity	Speci- ficity	False POS	False NEG
0.280	172	0	94	0	64.7	100.0	0.0	35.3	.
0.300	162	21	73	10	68.8	94.2	22.3	31.1	32.3
0.320	162	21	73	10	68.8	94.2	22.3	31.1	32.3
0.340	162	21	73	10	68.8	94.2	22.3	31.1	32.3
0.360	162	21	73	10	68.8	94.2	22.3	31.1	32.3
0.380	162	21	73	10	68.8	94.2	22.3	31.1	32.3
0.400	145	34	60	27	67.3	84.3	36.2	29.3	44.3
0.420	145	34	60	27	67.3	84.3	36.2	29.3	44.3
0.440	145	34	60	27	67.3	84.3	36.2	29.3	44.3
0.460	145	34	60	27	67.3	84.3	36.2	29.3	44.3
0.480	133	53	41	39	69.9	77.3	56.4	23.6	42.4
0.500	133	53	41	39	69.9	77.3	56.4	23.6	42.4
0.520	133	53	41	39	69.9	77.3	56.4	23.6	42.4
0.540	133	53	41	39	69.9	77.3	56.4	23.6	42.4
0.560	133	53	41	39	69.9	77.3	56.4	23.6	42.4
0.580	133	53	41	39	69.9	77.3	56.4	23.6	42.4
0.600	126	73	21	46	74.8	73.3	77.7	14.3	38.7
0.620	126	73	21	46	74.8	73.3	77.7	14.3	38.7
0.640	126	73	21	46	74.8	73.3	77.7	14.3	38.7
0.660	126	73	21	46	74.8	73.3	77.7	14.3	38.7
0.680	126	73	21	46	74.8	73.3	77.7	14.3	38.7
0.700	126	73	21	46	74.8	73.3	77.7	14.3	38.7
0.720	126	73	21	46	74.8	73.3	77.7	14.3	38.7
0.740	103	76	18	69	67.3	59.9	80.9	14.9	47.6
0.760	81	84	10	91	62.0	47.1	89.4	11.0	52.0
0.780	81	84	10	91	62.0	47.1	89.4	11.0	52.0
0.800	81	84	10	91	62.0	47.1	89.4	11.0	52.0
0.820	81	84	10	91	62.0	47.1	89.4	11.0	52.0
0.840	52	84	10	120	51.1	30.2	89.4	16.1	58.8
0.860	52	86	8	120	51.9	30.2	91.5	13.3	58.3
0.880	52	86	8	120	51.9	30.2	91.5	13.3	58.3
0.900	0	94	0	172	35.3	0.0	100.0	.	64.7



Here is the R program file assay.R (/stat504/sites/onlinecourses.science.psu.edu.stat504/files/lesson06/assay.R) that

corresponds to the SAS program assay4.sas.

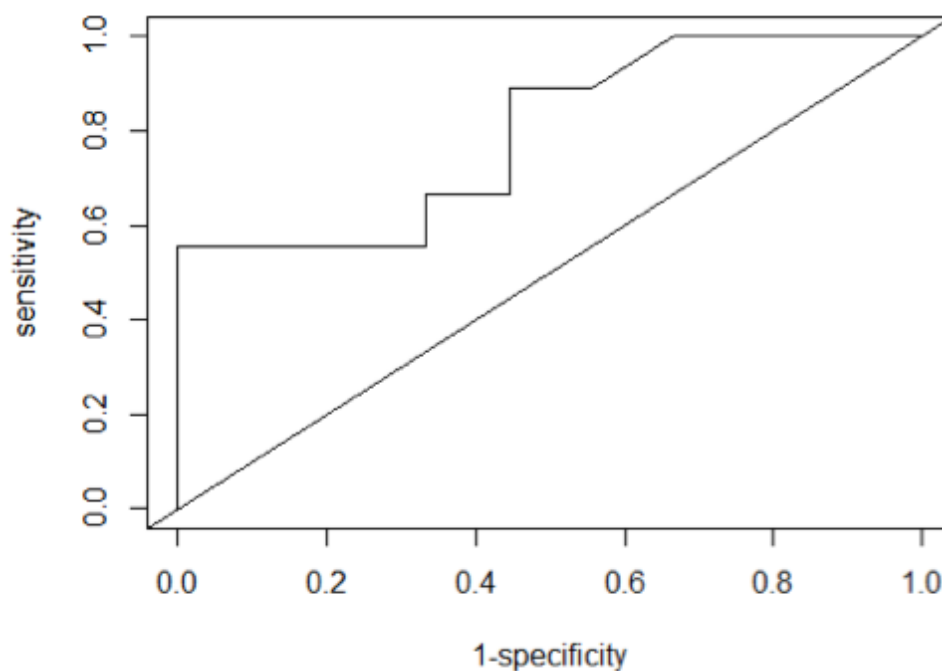
```
#### The following is a function adapted from
#### http://www.math.mcmaster.ca/peters4f03s4f03_0607index.html
#### roc.plot() will plot the ROC curve given two vectors of scores,
#### the first for the treatment group (y==1) and the second for the control group (y==0).

roc.plot <-
function (sd, sdc, newplot = TRUE, ...)
{
  sall <- sort(c(sd, sdc))
  sens <- 0
  specc <- 0
  for (i in length(sall):1) {
    sens <- c(sens, mean(sd >= sall[i], na.rm = T))
    specc <- c(specc, mean(sdc >= sall[i], na.rm = T))
  }
  if (newplot) {
    plot(specc, sens, xlim = c(0, 1), ylim = c(0, 1), type = "l",
         xlab = "1-specificity", ylab = "sensitivity", ...)
    abline(0, 1)
  }
  else lines(specc, sens, ...)
  npoints <- length(sens)
  area <- sum(0.5 * (sens[-1] + sens[-npoints]) * (specc[-1] -
    specc[-npoints]))
  lift <- (sens - specc)[-1]
  cutoff <- sall[lift == max(lift)][1]
  sensopt <- sens[-1][lift == max(lift)][1]
  specopt <- 1 - specc[-1][lift == max(lift)][1]
  list(area = area, cutoff = cutoff, sensopt = sensopt, specopt = specopt)
}

#### Let us draw the ROC plot

roc.plot(r,n-r)
```

Here is the ROC graph from R output:



The area under the curve is  $c = 0.746$  which indicates good predictive power of the model.

```
> roc.plot(r,n-r)
$area
[1] 0.7962963

$cuttoff
[1] 7

$sensopt
[1] 0.5555556

$specopt
[1] 1
```

---

◀ 7.2.2 - Overdispersion (/stat504/node/162)

up  
(/stat504/node/160)

7.3 - Binary Logistic Regression: Summary ▶  
(/stat504/node/221)

---