Journal of Statistical Software

November 2007, Volume 21, Issue 10. http://www.jstatsoft.org/

nonbinROC: Software for Evaluating Diagnostic Accuracies with Non-Binary Gold Standards

Paul Nguyen
University of Western Ontario

Abstract

ROC analysis is a standard method for estimating and comparing diagnostic tests' accuracies when the gold standard is binary. However, there are many situations when the gold standard is not binary. In these situations, traditional ROC methods applied have lead to biased and uninformative outcomes. This article introduces **nonbinROC**, software for R that implements nonparametric estimators proposed by Obuchowski (2005) for estimating and comparing diagnostic tests' accuracies when the gold standard is measured on a continuous, ordinal or nominal scale. The results produced from these estimators are interpreted in the same manner as in ROC analysis but are not associated with any ROC curve.

Keywords: ROC, gold standard, diagnostic test, continuous scale, ordinal scale, nominal scale.

1. Introduction

In medical research, the area under the receiver operating characteristic (ROC) curve is a standard measure for the evaluation of the accuracy of a diagnostic test. The ROC curve area is typically defined as the average value, which ranges from 0.5 to 1, of sensitivity for all possible values of specificity (Zhou, Obuchowski, and McClish 2002). In binary ROC analysis, there exists a gold or reference standard, which is independent of the diagnostic test, to indicate the true disease status of a patient. After the gold standard procedure and diagnostic test are performed on the patients, the diagnostic test results are compared with the gold standard results to estimate the accuracy at various cutpoints of the diagnostic test results (Zhou *et al.* 2002). However, the ROC curve area is equivalent to Harrell's c-index (Harrell Jr, Califf, Pryor, Lee, and Rosati 1982), which is defined as the probability that a randomly selected patient with the condition is ranked higher than a patient without the condition. This rank based measure can be calculated without constructing the ROC curve.

There are many situations in which the gold standard is measured on a non-binary scale. For example, the gold standard for acute abdominal pain in children can be measured as a categorical indicator of appendicitis, gastroenteritis, constipation, intestinal obstruction and urinary tract infection (Obuchowski, Goske, and Applegate 2001). A common approach for estimating the accuracy of a diagnostic test is to dichotomize the non-binary scale gold standard results, and then apply traditional ROC methods. However, this approach often conceals important relationships between the gold standard and diagnostic test and computes a biased estimate of the accuracy (Obuchowski 2005). There exist other approaches for dealing with these situations, such as applying a regression modelling framework to the ROC curve (Pepe 2000); however, this paper will only discuss the work researched by Obuchowski (2005, 2006) and Obuchowski, Goske, and Applegate (2001). Their work has proposed nonparametric estimators, which are based on a linear function of Kendall's τ (Hanley and McNeil 1982; Bamber 1975) and an extension of the Wilcoxon-Mann-Whitney estimate of accuracy (Snedecor and Cochran 1989), for calculating the accuracy of a diagnostic test when the gold standard is measured on a continuous, ordinal or nominal scale.

The objective of this paper is to introduce the R (R Development Core Team 2007) package, **nonbinROC**, which implements Obuchowski's methods. It is available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/`.

2. Background

We start with a brief review of Obuchowski's methods. The estimators for the accuracy of a diagnostic test when the gold standard is measured on a binary, continuous, ordinal or nominal scale are provided in Table 1. The accuracy estimates for the continuous, ordinal or nominal gold standards can be interpreted similarly in ROC analysis, but they are not associated with ROC curves (Harrell Jr *et al.* 1982).

The estimators for the continuous, ordinal or nominal gold standards are nonparametric, and hence, no assumptions on the distribution of the diagnostic tests and gold standard are needed.

Gold standard	Estimator of accuracy	
Binary	$\hat{\theta} = \frac{1}{n_t n_s} \sum_{i=1}^{n_t} \sum_{j=1}^{n_s} \Psi(X_{it}, X_{js})$	$\Psi = 1$ if $X_{it} > X_{js}$ $\Psi = 0.5$ if $X_{it} = X_{js}$ $\Psi = 0$ if $X_{it} < X_{js}$
Continuous	$\hat{\theta}' = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \Psi(X_{it}, X_{js})$	$\Psi = 1$ if $t > s$ and $X_{it} > X_{js}$ $\Psi = 1$ if $s > t$ and $X_{js} > X_{it}$ $\Psi = 0.5$ if $t = s$ or $X_{it} = X_{js}$ $\Psi = 0$ if otherwise
Ordinal	$\hat{\theta}'' = 1 - \sum_{t=1}^{T} \sum_{s>t}^{T} w_{ts} L(t,s)(1 - \hat{\theta}_{ts})$	$\hat{\theta}_{ts}$ is the binary scale estimator
Nominal	same as ordinal	$\hat{\theta}_{ts}$ is the binary scale estimator for $D_{(t-s)tj}$

Table 1: Estimators for diagnostic test accuracy.

Suppose the gold standard is measured on a binary scale with the two outcomes, t and s, being the true positive and true negative of the disease status, respectively. X_{it} is defined as the diagnostic test result for the ith patient with status t, and X_{js} is defined as the diagnostic test result for the jth patient with status s. The value of X is a subjective confidence score or an objective measurement, and n_t and n_s are the number of patients with states, t and s, in the sample, respectively. $\hat{\theta}$ is interpreted as the probability that if a patient is randomly selected from each group the true positive patient will score higher than the true negative patient.

Suppose the gold standard is measured on a continuous scale. Analogous to the binary scale gold standard, X_{it} is defined as the diagnostic test result for the ith patient who has a continuous scale gold standard outcome t. The value of X is a subjective confidence score or an objective measurement, and N is the total number of patients in the sample. $\hat{\theta}'$ estimates the probability that the ordering of the diagnostic tests will match the true ordering of the patients.

Suppose the gold standard is measured on an ordinal scale with T total number of disease states or categories. X_{it} is defined as the diagnostic test result for the ith patient who has a gold standard outcome $t = 1, 2, \ldots, T$. w_{ts} is defined as a weight for the states, t and s, such that $w_{ts} = n_t n_s \Big/ \sum_{i=1}^{T} \sum_{j>i}^{T} n_i n_j$. $L(t, s)$ is a loss or penalty function, with a value between 0 and 1, for misclassifying patients between states, t and s, for $s > t$. $L(t, s) = 1$ indicates the greatest penalty and $L(t, s) = 0$ indicates no penalty. The common value for all entries of L is 1, but some people may assign less penalty for neighboring states. For example, given 4 states, a possible scheme for L is $L(t, t + 1) = 0.25$, $L(t, t + 2) = 0.5$ and $L(t, t + 3) = 1$.

For the ordinal scale gold standard, the estimates of pairwise accuracy, $\hat{\theta}_{ts}$, and the summary or overall accuracy, $\hat{\theta}''$, are both important. $\hat{\theta}_{ts}$ is defined as the binary scale estimator of the diagnostic test accuracy for distinguishing between the states, t and s. It has the same interpretation as $\hat{\theta}$. There are $T \times (T - 1)/2$ estimates of pairwise accuracy. Given the weighing scheme, w, and the defined penalty, L, $\hat{\theta}''$ is the probability that if a patient is randomly selected from each state the patient in the higher state will score higher than the patient in the lower state.

Suppose the gold standard is measured on a nominal scale with T total number of disease states or categories. Most of the notation for the nominal scale gold standard is the same with the ordinal scale gold standard. Here, X is a $(1 \times T)$ vector of confidence scores, one score for each of the T states. For example, given 3 states, if the patient shows symptoms from state 2, this patient may be assigned a confidence score of 5% for state 1, 90% for state 2 and 5% for state 3. The sum of the confidence scores should sum to 1 or 100% for each patient. $D_{(t-s)tj}$, which is used to calculate $\hat{\theta}_{ts}$, is defined as the difference in confidence scores assigned to states t and s $(s > t)$ for the jth patient with disease status t. $\hat{\theta}_{ts}$ is interpreted in the same manner as the ordinal scale gold standard but $\hat{\theta}''$ is interpreted differently. Given the weighing scheme, w, and the defined penalty, L, $\hat{\theta}''$ is the probability that if a patient is selected from each state the patients are correctly ranked.

3. Package nonbinROC in use

The R (R Development Core Team 2007) package, **nonbinROC**, contains nonparametric statistical methods for estimating and comparing accuracies of diagnostic tests when the gold

Functions	
`contROC`	Computes the accuracy of a test and compares the accuracies of competing tests for a continuous scale gold standard.
`ordROC`	Computes the accuracy of a test and compares the accuracies of competing tests for a ordinal scale gold standard.
`nomROC`	Computes the accuracy of a test and compares the accuracies of competing tests for a nominal scale gold standard.
Datasets	
`tumor`	Contains continuous scale measurements of the renal tumor mass size for 74 patients.
`blood`	Contains continuous scale measurements of the blood iron concentration for 55 anaemia female patients.
`heart`	Contains ordinal measurements of magnetic resonance imaging and positron emission tomography scans for the heart tissues for 241 fictitious patients after myocardial infarction.
`abpain`	Contains confidence scores of the pre-imaging and post-imaging diagnoses for 60 patients suffering from acute abdominal pain.

Table 2: Summary of functions and datasets in the package.

standard is measured on a continuous, ordinal or nominal scale. Examples from Obuchowski (2005, 2006) and Obuchowski, Goske, and Applegate (2001) are also included in this package. Table 2 provides a summary of the objects in this package.

3.1. Continuous scale gold standard

The function `contROC()` computes the accuracy of the diagnostic test and compares the accuracies of competing diagnostic tests when the gold standard is measured on a continuous scale.

```
contROC(gldstd, test1, test2 = NULL)
```

It requires the two arguments, `gldstd` and `test1`, for the results in vector form of the continuous scale gold standard and diagnostic test, respectively. The last argument, `test2`, is for the results of another optional diagnostic test.

This function returns the estimate and standard error for the accuracy of a diagnostic test compared to the continuous scale gold standard. If two diagnostic tests are presented, the covariance and a test statistic are also returned. The test statistic is for a two-sided alternative hypothesis and is compared to the standard Normal distribution.

In this package, the `tumor` dataset, provided by Obuchowski (2005), and the `blood` dataset, provided by Obuchowski (2006), have continuous scale gold standards. A demonstration will be given on the `tumor` dataset. This dataset contains a series of continuous scale measurements (in cm) for the size of the renal tumor mass for 74 patients based on surgery (`SURG`), a computed tomography (`CT`) and a fictitious test (`Fi`).

```
  SURG  CT  Fi
1  3.3 3.9 3.0
2  1.9 2.0 2.2
3  4.0 3.7 4.1
4  3.5 3.1 3.6
5  3.0 3.0 2.9
...
```

The set of measurements based on surgery is the gold standard and the remaining two sets of measurements are the two competing diagnostic tests. The fictitious test was created to illustrate the proposed method for comparing the accuracies of two diagnostic tests in a paired design (Obuchowski 2005). The analysis for this example is provided below:

```
R> data("tumor")
R> attach(tumor)
R> contROC(SURG, CT, Fi)

$Accuracy
  Estimates Standard.Errors
1 0.8709737     0.020975482
2 0.9563125     0.007080425

$Covariance
[1] 4.904536e-05

$`Two-Sided Hypothesis Test`
          Z      p.value
1 -4.310189 1.631148e-05
```

This output is similar to the values provided by Obuchowski (2005) in which the estimated accuracies of the computed tomography and fictitious test are 0.871 and 0.957, respectively. The results are interpreted by the following:

- Of two randomly chosen renal masses, there is a 87.1% chance that the larger renal mass (as determined at surgery) will have a larger measured diameter on computed tomography than the smaller renal mass.
- Of two randomly chosen renal masses, there is a 95.7% chance that the larger renal mass (as determined at surgery) will have a larger measured diameter on the fictitious test than the smaller renal mass.
- There is statistical evidence that the fictitious discriminates between the renal masses of different sizes better than computed tomography.

3.2. Ordinal scale gold standard

The function `ordROC()` computes the accuracy of the diagnostic test and compares the accuracies of competing diagnostic tests when the gold standard is measured on an ordinal scale.

```
ordROC(gldstd, test1, test2 = NULL, penalty = NULL)
```

It requires the two arguments, `gldstd` and `test1`, for the results in vector form of the ordinal scale gold standard and diagnostic test, respectively. The argument, `test2`, is for the results of another optional diagnostic test. The last argument, `penalty`, is for penalty function matrix $L[i, j]$ in which $0 \leq L[i, j] \leq 1$ for $j > i$.

This function returns the estimates and standard errors for all pairs of categories of the ordinal scale gold standard. With these values and the penalty function matrix, which is also returned, the estimate and standard error for the summary accuracy of a diagnostic test compared to the gold standard are computed and returned. If two diagnostic tests are presented, the covariance of the summary accuracy and a test statistic are also returned. The test statistic is for a two-sided alternative hypothesis and is compared to the standard Normal distribution.

An example of this function is demonstrated on the `heart` dataset, provided by Obuchowski (2005), located in this package. This dataset contains a series of ordinal measurements of the positron emission tomography (`PET`) and magnetic resonance imaging (`MRI`) scans for the heart tissues of 241 fictitious patients after experiencing myocardial infarctions. The ordinal measurements based on the positron emission tomography evaluate the tissue in the most damage part of the heart with the values: 1 = normal, 2 = ischemic, 3 = hibernating and 4 = necrotic. The ordinal measurements based on the magnetic resonance imaging evaluate amount of scarring present in the most damaged segment of the heart with the values: 0 = normal, 1 = 1–24%, 2 = 25–49%, 3 = 50–74%, 4 = 75–99% and 5 = 100%.

```
  PET MRI
1   1   0
2   1   0
3   1   0
4   1   0
5   1   0
...
```

The set of measurements based on positron emission tomography is the gold standard and the other set of measurements based on magnetic resonance imaging is the diagnostic test. The analysis for this example is presented below:

```
R> data("heart")
R> attach(heart)
R> penalty <- matrix(c(0,0,0,0,0.25,0,0,0,0.5,0.25,0,0,1,0.5,0.25,0),
+   nrow = 4)
R> ordROC(PET, MRI, penalty = penalty)

$`Pairwise Accuracy`
    Pair  Estimate Standard.Error
1 1 vs 2 0.5267335     0.06621658
2 1 vs 3 0.8072484     0.05023689
3 1 vs 4 0.7699133     0.03445891
```

```
4 2 vs 3 0.7857143     0.06888356
5 2 vs 4 0.7520525     0.04986680
6 3 vs 4 0.5317604     0.06269441

$`Penalty Matrix`
  1    2    3    4
1 0 0.25 0.50 1.00
2 0 0.00 0.25 0.50
3 0 0.00 0.00 0.25
4 0 0.00 0.00 0.00

$`Overall Accuracy`
  Estimate Standard.Error
1 0.824842     0.02170501
```

The output agrees with the values provided by Obuchowski (2005) in which the estimated overall accuracy of the magnetic resonance imaging is 0.825 for the given penalty function. Thus, of two randomly chosen patients with different heart muscle damage with the given penalty function, the magnetic resonance imaging has a 82.5% chance of revealing more scar in the patient with more tissue damage.

3.3. Nominal scale gold standard

The function `nomROC()` computes the accuracy of the diagnostic test and compares the accuracies of competing diagnostic tests when the gold standard is measured on a nominal scale. This function has a similar structure to the `ordROC()` function.

```
nomROC(gldstd, test1, test2 = NULL, penalty = NULL)
```

The arguments of `nomROC()` are previously defined by the arguments of `ordROC()`. The only difference between `nomROC()` and `ordROC()` is the input for the arguments, `test1` and `test2`, for `nomROC()` is in a data frame or matrix form.

An example of this function is demonstrated on the `abpain` dataset, provided by Obuchowski, Goske, and Applegate (2001), located in this package. This dataset contains a series of confidence scores for the pre-imaging (`Pre1`, `Pre2` and `Pre3`) and post-imaging (`Post1`, `Post2` and `Post3`) diagnoses of 60 patients suffering from acute abdominal pain (`Group`) grouped into the following 3 distinct categories: 1 = surgical abdominal or urogenital condition, 2 = non-surgical abdominal condition and 3 = non-surgical urogenital condition.

```
  Group Pre1 Pre2 Pre3 Post1 Post2 Post3
1     1   75   25    0   100     0     0
2     1   15   85    0   100     0     0
3     1    5   95    0   100     0     0
4     1  100    0    0   100     0     0
5     1   40   60    0   100     0     0
...
```

These 3 categories are considered as the nominal scale gold standard and the two set of confidence scores for pre-imaging and post-imaging diagnoses are the diagnostic tests. The analysis for this example is presented below:

```
R> data("abpain")
R> attach(abpain)
R> pre <- data.frame(Pre1, Pre2, Pre3)
R> post <- data.frame(Post1, Post2, Post3)
R> penalty <- matrix(c(0,0,0,1,0,0,1,0.5,0), nrow = 3)
R> nomROC(Group, pre, post, penalty)

$`Pairwise Accuracy for Test 1`
    Pair  Estimate Standard.Error
1 1 vs 2 0.6291667     0.08657834
2 1 vs 3 0.8549107     0.07025674
3 2 vs 3 0.8761905     0.07224316

$`Pairwise Accuracy for Test 2`
    Pair  Estimate Standard.Error
1 1 vs 2 0.9614583     0.02546630
2 1 vs 3 0.9732143     0.02812582
3 2 vs 3 0.8166667     0.07995460

$`Penalty Matrix`
  1 2   3
1 0 1 1.0
2 0 0 0.5
3 0 0 0.0

$`Overall Accuracy`
   Estimate Standard.Error
1 0.7895907     0.04626797
2 0.9439502     0.02182565

$Covariance
[1] 0.0004802158


$`Two-Sided Hypothesis Test`
          Z      p.value
1 -3.792430 0.0001491806
```

The output agrees with the values provided by Obuchowski, Goske, and Applegate (2001) in which the estimated overall accuracy of the pre-imaging and post-imaging diagnoses are 0.790 and 0.944, respectively. The results are interpreted by the following:

- Given penalty function, there is a 79.0% chance that physicians correctly triage the patients prior to diagnostic imaging.

- Given penalty function, there is a 94.4% chance that physicians correctly triage the patients after diagnostic imaging.
- There is statistical evidence that imaging improves the overall accuracy of physicians to correctly triage the patients.

4. Summary

nonbinROC is an R package for estimating and comparing the accuracies of diagnostic tests when the gold standard is measured on a continuous, ordinal or nominal scale. It uses methods proposed by Obuchowski (Obuchowski 2005, 2006; Obuchowski *et al.* 2001). The methods are nonparametric and make no assumptions about the distribution of the diagnostic test and gold standard results. The accuracy estimates produced from these methods are evaluated in the same manner as in ROC analysis, but they are not associated with ROC curves.

Acknowledgments

The author thanks Duncan Murdoch and Linh P. Nguyen for their comments on the software and article.

References

Bamber D (1975). "The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Characteristic Graph." *Journal of Mathematical Psychology*, **12**, 387–415.

Hanley JA, McNeil BJ (1982). "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve." *Radiology*, **143**, 29–36.

Harrell Jr FE, Califf RM, Pryor DB, Lee KL, Rosati RA (1982). "Evaluating the Yield of Medical Tests." *Journal of the American Medical Association*, **247**, 2543–2546.

Obuchowski NA (2005). "Estimating and Comparing Diagnostic Tests' Accuracy When the Gold Standard Is Not Binary." *Academic Radiology*, **12**, 1198–1204.

Obuchowski NA (2006). "An ROC-Type Measure of Diagnostic Accuracy When the Gold Standard is Continuous-Scale." *Statistics in Medicine*, **25**, 481–493.

Obuchowski NA, Goske MJ, Applegate KE (2001). "Assessing Physician's Accuracy in Diagnosing Paediatric Patients with Acute Abdominal Pain: Measuring Accuracy for Multiple Disease." *Statistics in Medicine*, **20**, 3261–3278.

Pepe MS (2000). "Receiver Operating Characteristic Methodology." *Journal of the American Statistical Association*, **95**, 308–311.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Snedecor GW, Cochran WG (1989). *Statistical Methods*. 8th edition. Iowa State University Press, Ames, Iowa.

Zhou XH, Obuchowski NA, McClish DK (2002). *Statistical Methods in Diagnostic Medicine*. Wiley-Interscience, New York, NY.

Affiliation:

Paul Nguyen
Department of Statistical and Actuarial Sciences
University of Western Ontario
London, Ontario, Canada
E-mail: pnguye3@uwo.ca

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 21, Issue 10 *Submitted:* 2006-09-25
November 2007 *Accepted:* 2007-09-06