ROC analysis of clustered data with R

Description

The R script `funcs_clusteredROC.R` contains functions to perform the statistical methods in:

Obuchowski NA. Nonparametric analysis of clustered ROC curve data. *Biometrics*. 1997: 567-578.

The main function called by the user is `clusteredROC()`. It can handle data for a single ROC curve or for two correlated ROC curves. The response cannot be missing – an error will result. If the predictor for either curve is missing, the entire record is removed. Below is a table of the arguments for `clusteredROC()`.

Argument	Description	MRA example from Obuchowski 1997
predictor1	a vector containing the predictor for ROC curve 1	the test result for an artery from Reader 1
predictor2	a vector containing the predictor for ROC curve 2 (this can be omitted if you are only estimating 1 ROC curve)	the test result for an artery from Reader 2
response	a vector containing the response for both ROC curves	the true disease state for an artery (1=significant disease, 0=not significant disease)
clusterID	a vector containing IDs for the clusters	patient ID
alpha	the type I error rate	0.05
level	can be used to specify the response level considered ‘positive’ (if omitted, the second level of the response is selected)	‘1’ (significant disease) is considered positive
print.all	if TRUE, intermediate estimates are printed	

MRA example: 2 correlated ROC curves

Here we reproduce the results from the illustrative example in Obuchowski 1997. The data from Table 3 is contained in the file `MRA.csv`. Read in this data and the functions provided in `funcs_clusteredROC.R`.

```r
df = read.csv("H:\\MRA.csv")
source("H:\\funcs_clusteredROC.R")
head(df)
```

```
##   patient_id artery_side reader1_result reader2_result disease
## 1          1        left             87             87       1
## 2          1       right             79             83       1
## 3          2        left             88             94       1
## 4          2       right             95             93       1
## 5          3        left            100            100       1
## 6          3       right             68             79       1
```

You can use the following code to compare the accuracy of the two readers.

```
clusteredROC(predictor1 = df$reader1_result,
             predictor2 = df$reader2_result,
             response   = df$disease,
             clusterID  = df$patient_id,
             alpha      = 0.05)
```

```
##
## Total # of clusters: 36
## Total # of observations: 65
## Min # of observations per cluster: 1
## Max # of observations per cluster: 2
## AUC (SE) for ROC curve 1: 0.9837 (0.0108)
## AUC (SE) for ROC curve 2: 0.9852 (0.0097)
## Difference (SE): 0.0014 (0.0066)
## 95% CI for difference: (-0.0115, 0.0143)
## Associated p-value: 0.8271
```

There are 65 arteries from 36 patients in this sample. As reported in Obuchowski 1997 on page 574, the estimated area at the artery level is 0.984 (SE = 0.011) for Reader 1 and 0.985 (SE = 0.010) for Reader 2. The estimated difference between these areas is 0.001 (SE = 0.007). Note that the p-value and 95% confidence interval from `clusteredROC()` are slightly different from those reported in Obuchowski 1997. This is because rounded values were used for the difference and standard error in Obuchowski 1997 for illustrative purposes. We can check that all intermediate estimates from `clusteredROC()` align with those presented in the footer of Table 3 from Obuchowski 1997 by setting `print.all = TRUE`.

```
clusteredROC(predictor1 = df$reader1_result,
             predictor2 = df$reader2_result,
             response   = df$disease,
             clusterID  = df$patient_id,
             alpha      = 0.05,
             print.all  = TRUE)
```

```
##
## Total # of clusters: 36
## Total # of observations: 65
## Min # of observations per cluster: 1
## Max # of observations per cluster: 2
## AUC (SE) for ROC curve 1: 0.9837 (0.0108)
## AUC (SE) for ROC curve 2: 0.9852 (0.0097)
## Difference (SE): 0.0014 (0.0066)
## 95% CI for difference: (-0.0115, 0.0143)
## Associated p-value: 0.8271
##
##            name         value
## 1             I 36.0000000000
## 2           I10 23.0000000000
## 3           I01 27.0000000000
## 4             M 29.0000000000
## 5             N 36.0000000000
## 6  reader 1 S10  0.0013151307
## 7  reader 1 S01  0.0022375873
## 8  reader 1 S11  0.0051785478
## 9  reader 2 S10  0.0009284122
## 10 reader 2 S01  0.0022598583
## 11 reader 2 S11 -0.0005015760
## 12       S10_12  0.0008484357
## 13       S01_12  0.0019241250
## 14       S11_12  0.0028648340
## 15       S11_21 -0.0015136931
```

MRA example: 1 ROC curve

To analyze a single ROC curve, simply omit the predictor2 argument. Here we look at the area under the ROC curve just for Reader 1.

```
clusteredROC(predictor1 = df$reader1_result,
             response   = df$disease,
             clusterID  = df$patient_id,
             alpha      = 0.05)
```

```
##
## Total # of clusters: 36
## Total # of observations: 65
## Min # of observations per cluster: 1
## Max # of observations per cluster: 2
## AUC (SE) for ROC curve: 0.9837 (0.0108)
## 95% CI for AUC: (0.9625, 1.005)
```

Again, there are 65 arteries from 36 patients. The results displayed are the same as before, except now a 95% confidence interval for the area under the ROC curve for Reader 1 is provided.