Regression Analysis for Correlated ROC Curves

Krisztian Sebestyen and Xiao-Hua Zhou

January 13, 2011

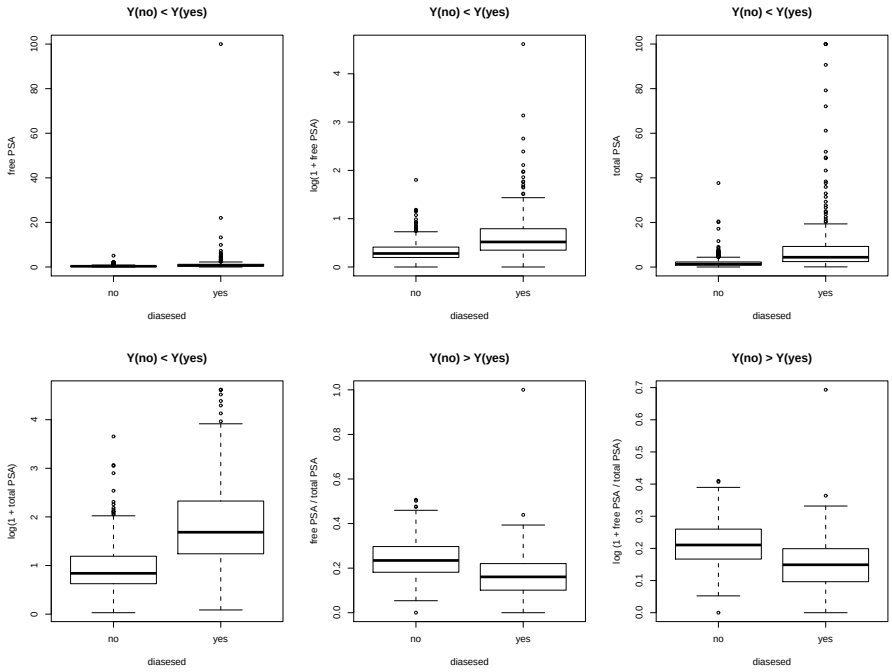

Figure 1: Boxplots of free, total and ratio of free to total PSA in non- and diseased groups. Note that ratio PSA appears to be higher in the non-diseased groups

Introduction

This document describes the implementation and illustrations of ROC regression models for continuous scale test results, described in Chapter 9 of the book. The code was written by Krisztian Sebestyen. The dataset is a subset of the CARET dataset and can be downloaded from http://labs.fhcrc.org/pepe/book/data/psa2b.csv. In the study the objective was to model receiver operating characteristic (ROC) curves for Prostate-specific antigen (PSA) over time where PSA measurements were taken at *irregular* times prior to diagnosis of prostate cancer. Below we describe implementation of the correlated ROC regression approach described in Section 9.1.3 of the book. This approach is termed 'Direct Continuous Correlated ROC' and is based on an estimating equation approach. The corresponding R-command is 'dccorroc'.

The boxplot displays free (fpsa), total (tpsa) and the ratio of free to total PSA (rpsa) in the nondiseased and diseased groups. All outcomes were transformed to approximate symmetry via the $log(1+x)$. Note that rpsa is higher in the nondisased group therefore *both rpsa and logrpsa were multiplied by* -1 *for the analyses below.*

Below is an excerpt of the dataset. We generated the variable 'occasion' that denotes repeated measurements on a subject over irregular time points. There are at most 9 measurements on a subject. Note that when comparing the diagnostic accuracy of two markers/tests, say total PSA and ratio of free to total PSA, there are *two sources of correlation.* One is due to a given marker measured on the same subject over time, denoted by the variable 'id', the other is due to the two markers measured on a patient at a given time. *When comparing the diagnostic accuracy of tpsa and rpsa via 'dccorroc', we ignore the correlation due to longitudinal measurements and pool the data over 6 occasions. Thus the only correlation accounted for by 'dccorroc' is the correlation of two tests measured on the same subject at a given time* ($Q = 2$).

id	d	t	fpsa	tpsa	age	occasion	qage	rpsa	logtpsa	logfpsa	logrpsa
1	1	-4.48	3.52	14.82	67.58	1	1	0.24	2.76	1.51	0.21
2	1	-4.50	1.10	5.54	70.17	1	1	0.20	1.88	0.74	0.18
2	1	-1.34	2.40	8.15	73.33	2	1	0.30	2.21	1.23	0.26
2	1	-0.36	2.43	10.71	74.31	3	1	0.23	2.46	1.23	0.20
3	0	-3.38	0.23	0.94	55.03	1	0	0.24	0.66	0.20	0.22
3	0	-1.12	0.23	1.03	57.29	2	0	0.22	0.71	0.20	0.20
3	0	-0.18	0.24	1.03	58.23	3	0	0.23	0.71	0.22	0.21
3	0	0.84	0.20	0.98	59.25	4	0	0.21	0.68	0.18	0.19
3	0	1.85	0.15	0.78	60.26	5	0	0.19	0.58	0.14	0.17
3	0	2.85	0.32	1.23	61.26	6	0	0.26	0.80	0.28	0.23
3	0	3.80	0.34	1.51	62.21	7	0	0.23	0.92	0.29	0.20
3	0	4.78	0.31	1.46	63.19	8	0	0.21	0.90	0.27	0.19
3	0	5.78	0.40	1.19	64.19	9	0	0.33	0.78	0.34	0.29

Correlated ROC Regression Models

Let X be a vector of covariates common to both patients with and without the condition and X_D be a vector of covariates specific to the diseased population. Let T_{1q} and T_{0q} be continuous-scale test (marker) results measured at the q^{th} ($q = 1..Q$) occasion for a patient in the diseased and nondiseased populations, respectively. If $\bar{F}_{1q,x,x_D}$ and $\bar{F}_{0q,x}$ denote the covariate specific survival functions of the diseased and non-diseased populations, then the ROC regression that accounts for correlation that arises for Q measurements taken on the subject at false positive rate (FPR) $p \in (0, 1)$ is of the form:

$$\begin{aligned} ROC_{q,x,x_D}(p) &= \bar{F}_{1q,x,x_D}\left(\bar{F}^{-1}_{0q,x}(p)\right) && (1) \\ &= g_q[\gamma' h(p) + \beta' x + \beta' x_D] = g_q\{\theta'[h(p)', x', x_D']'\} && (2) \end{aligned}$$

where $\theta = (\gamma', \beta', \beta_D')'$ and g_q and $h(p)$ are known link and basis functions at occasion q. For more details see (2).

R command

The R command to run the regression is 'dccorroc'. The estimating equations use 'dfsane()' of package 'BB'. Package 'abind' is also required to combine multidimensional arrays after bootstrapping.

```
x <- cbind(expand.grid(c(0,-2,-4,-8)));colnames(x) <- c("t")
pepe.11_6b <- dccorroc(
theta = rep(c(-1,1,0),2),
data = data[data$occasion <= 6,],
tests = ~ logtpsa + logrpsa ,
formula = ~1,
formulaD = ~t,
diseased = ~d,
covariates = x,
fpr = (1:50)/51,
return.covariance = T,NBOOT = 20,SEED = 1:20,
trace = 0,
control = list(tol = 1e-5,trace = T,maxit = 200))

## plot
require(graphics)
postscript("table1.EtzioniPepe.ps")
plot.dccorroc(pepe.11_6b,x=x,main = 'ROC curves corresponding to table 1
(Etzioni-Pepe-Hu-Goodman)')
dev.off()

## table
sm <- summary.dccorroc(pepe.11_6b)
a0.t <- outer( pepe.11_6b$theta["t",] , c(0,-2,-4,-8)) + pepe.11_6b$theta[1,]
table1.EtzioniPepe <- list(
 logtpsa = cbind(a0.t = a0.t["logtpsa",],
 a1.t = pepe.11_6b$theta["slope","logtpsa"],
 auc.t = sm$auc[,"Estimate","logtpsa"]),
 neglogrpsa = cbind(a0.t = a0.t["logrpsa",],
 a1.t = pepe.11_6b$theta["slope","logrpsa"],
 auc.t = sm$auc[,"Estimate","logrpsa"])
)
require(Hmisc)
z <- signif(do.call('rbind',table1.EtzioniPepe),3)
rownames(z) <- rep(paste("time=",-c(0,-2,-4,-8),sep=''),2)
colnames(z) <- c("a0(t)","a1(t)","auc(t)")
latex(z,file = "table1.EtzioniPepe.tex",title = '',n.rgroup = c(4,4),
rgroup = c("total PSA","ratio PSA"))
```

The summary object returned after rounding is:

```
$covariates
        t
time=0  0
time=2 -2
time=4 -4
time=8 -8
attr(,"assign")
[1] 1

$auc
, , logtpsa

       Estimate Std. Error z value Pr(>|z|)
time=0     0.93         NA      NA       NA
time=2     0.89         NA      NA       NA
time=4     0.82         NA      NA       NA
time=8     0.65         NA      NA       NA

, , logrpsa

       Estimate Std. Error z value Pr(>|z|)
time=0     0.80         NA      NA       NA
time=2     0.77         NA      NA       NA
time=4     0.73         NA      NA       NA
time=8     0.64         NA      NA       NA


$dauc
, , 1-2

       Estimate Std. Error z value Pr(>|z|)
time=0     0.13       0.02    5.50     0.00
time=2     0.12       0.02    6.14     0.00
time=4     0.10       0.02    4.95     0.00
time=8     0.01       0.04    0.19     0.85


$coefficients
, , logtpsa

            Estimate Std. Error z value Pr(>|z|) 0.025 0.975
(Intercept)    -1.98       0.18  -10.87        0 -2.33 -1.62
slope           0.90       0.10    9.25        0  0.71  1.08
t              -0.18       0.03   -5.42        0 -0.25 -0.12

, , logrpsa
```

```
            Estimate Std. Error z value Pr(>|z|) 0.025 0.975
(Intercept)    -1.06       0.11   -9.79        0 -1.27 -0.85
slope           0.77       0.06   13.25        0  0.65  0.88
t              -0.07       0.02   -3.27        0 -0.12 -0.03


$dcoefficients
, , 1-2

            Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.92       0.18   -5.08     0.00
slope           0.13       0.09    1.36     0.17
t              -0.11       0.03   -3.22     0.00
```

Table 1: AUC's of correlated tests logtpsa and logrpsa (Q = 2) pooled over 6 irregular time points. This table closely matches table 1 of (1)

	a0(t)	a1(t)	auc(t)
total PSA			
time=0	−1.980	0.895	0.930
time=2	−1.610	0.895	0.885
time=4	−1.250	0.895	0.824
time=8	−0.526	0.895	0.653
ratio PSA			
time=0	−1.060	0.767	0.800
time=2	−0.912	0.767	0.765
time=4	−0.764	0.767	0.728
time=8	−0.467	0.767	0.645

results

Figure 2: Correlated ROC Regression approach