survBootOutliers: An R package for outlier detection in survival analysis

Jo<U+00E3>o Diogo Pinto, joao.pinto@tecnico.ulisboa.pt

April 19, 2017

Contents

1 Introduction

This package provides three new outlier detection methods to perform outlier detection in a survival analysis context. The first method OSD, for One Step Deletion, is a sequential procedure that maximizes the c-index of a fitted Cox regression using a a greedy one-step-ahead search, in each step the observation that when removed maximizes the concordance increase is permanently deleted from the dataset, the algorithm ends until k observations are removed, these are considered the most outlying ones. The second and third methods are based on bootstrap methods. The second method BHT, for Bootstrap Hypothesis Testing, is based on creating B bootstrap samples for each observation that is removed from the dataset, then an hypothesis test is made for the B concordance variations to be larger than zero, the observations with the lowest p-values are considered the most outlying. The last method DBHT, for Dual Bootstrap Hypothesis Testing, draws 2B bootstrap samples for each observation, B samples with each observation absent, just like with BHT, the other B bootstrap samples are drawn with the observation under test being deliberately inserted in each of the bootstrap samples. The hypothesis test is different, the two histograms are tested for inequality, for non-outlying observations the histograms are expected to be similar but for outlying observtions the histograms drawn when the observation is absent is expected to have higher concordance on average.

The package still provides three other methods considered more traditional based on Martingale-based residuals, Deviancre resiudals and Cox likelihood displacement.

These methods are based on the Master Thesis at Instituto Superior T<U+00E9>cnico, named ”Outlier detection in survival analysis” evaluated in May 2015. The link for the full text is left here for more detail: .

2 Example data

The well-known Worcester Heart Attack Study data is given as example and provided within the package;

```
> library(survBootOutliers)
> whas100_data <-  get.whas100.dataset()
> summary(whas100_data)
```

```
       X               los              age             gender    
 Min.   :  1.00   Min.   : 1.00   Min.   :32.00   Min.   :0.00  
 1st Qu.: 25.75   1st Qu.: 4.00   1st Qu.:59.75   1st Qu.:0.00  
 Median : 50.50   Median : 5.00   Median :71.00   Median :0.00  
 Mean   : 50.50   Mean   : 6.84   Mean   :68.25   Mean   :0.35  
 3rd Qu.: 75.25   3rd Qu.: 7.00   3rd Qu.:80.25   3rd Qu.:1.00  
 Max.   :100.00   Max.   :56.00   Max.   :92.00   Max.   :1.00  
      bmi            times          status    
 Min.   :14.92   Min.   :   6   Min.   :0.00  
 1st Qu.:23.54   1st Qu.: 715   1st Qu.:0.00  
 Median :27.19   Median :1878   Median :1.00  
 Mean   :27.04   Mean   :1505   Mean   :0.51  
 3rd Qu.:30.35   3rd Qu.:2076   3rd Qu.:1.00  
 Max.   :39.94   Max.   :2719   Max.   :1.00  
```

3 Examples

3.1 OSD

```
> whas <- get.whas100.dataset()
> outliers_osd <- survBootOutliers(
+     surv.object=Surv(time = whas$times,event = whas$status ),
+     covariate.data = whas[,2:5] , sod.method = "osd",
+     max.outliers = 10 )
> print(outliers_osd)
```

```
   removed_indexes
1                1
2               67
3               97
```

```
4               51
5               23
6               31
7               93
8               52
9               56
10              57
```

3.2 BHT

```
> whas <- get.whas100.dataset()
> outliers_bht <- survBootOutliers(
+     surv.object=Surv(time = whas$times,event = whas$status ),
+     covariate.data = whas[,2:5],
+     sod.method = "bht",
+     B = 100,
+     B.N = 100,
+     parallel.param = SnowParam() )
> print(outliers_bht)
```

```
      obs_id   avg_delta  max_delta pvalue
 [1,]     81 0.019496393 0.09001866   0.26
 [2,]      8 0.022594089 0.10069594   0.27
 [3,]     27 0.021328754 0.10469265   0.28
 [4,]     30 0.017967909 0.11092741   0.28
 [5,]     11 0.024924554 0.10359979   0.29
 [6,]      7 0.015321648 0.08078298   0.30
 [7,]     56 0.016622751 0.10277255   0.30
 [8,]     72 0.019719245 0.10414562   0.30
 [9,]     82 0.016312698 0.09350968   0.30
[10,]     90 0.015544647 0.13265076   0.30
[11,]      1 0.019655929 0.11260841   0.31
[12,]     28 0.018409953 0.10572962   0.31
[13,]     49 0.016814835 0.09205412   0.31
[14,]     45 0.015588261 0.09468507   0.32
[15,]     91 0.014148419 0.11526600   0.32
[16,]     46 0.013965082 0.08792967   0.33
[17,]     57 0.017565279 0.09308320   0.33
[18,]     93 0.019743702 0.11837145   0.33
[19,]      6 0.015371877 0.10806765   0.34
[20,]     25 0.011606613 0.12409427   0.34
[21,]     51 0.014801984 0.08563228   0.34
[22,]     58 0.015788058 0.11830485   0.34
[23,]     88 0.014222796 0.07744139   0.34
[24,]     17 0.015282910 0.12670356   0.35
[25,]     43 0.011645601 0.08009656   0.35
```

```
[26,]     44 0.017692737 0.10654893   0.35
[27,]     68 0.013369909 0.13306055   0.35
[28,]     78 0.016959295 0.11374909   0.35
[29,]      4 0.009874716 0.11385396   0.36
[30,]     13 0.018210961 0.10552769   0.36
[31,]     23 0.015736709 0.09409141   0.36
[32,]     26 0.011432656 0.11403087   0.36
[33,]     60 0.011303792 0.11584581   0.36
[34,]     74 0.012031085 0.11851099   0.36
[35,]     77 0.003289748 0.09969941   0.36
[36,]      2 0.012814310 0.10732768   0.37
[37,]     16 0.012315562 0.09986694   0.37
[38,]     31 0.016186459 0.10366624   0.37
[39,]     33 0.009901926 0.10261060   0.37
[40,]     48 0.010571140 0.10442507   0.37
[41,]     52 0.013907353 0.11045549   0.37
[42,]     67 0.016606840 0.10359273   0.37
[43,]     69 0.011573249 0.08456376   0.37
[44,]     73 0.013975527 0.11430791   0.37
[45,]     89 0.012553735 0.13129321   0.37
[46,]     97 0.018427862 0.10598816   0.37
[47,]     21 0.009943381 0.08576075   0.38
[48,]     32 0.010585528 0.10157520   0.38
[49,]     39 0.010318913 0.08835370   0.38
[50,]     42 0.009240086 0.09582555   0.38
[51,]     47 0.013393733 0.13690392   0.38
[52,]     71 0.007480619 0.10673378   0.38
[53,]     76 0.011626931 0.08685448   0.38
[54,]     84 0.008076102 0.09055220   0.38
[55,]    100 0.016466195 0.13526321   0.38
[56,]     12 0.010187662 0.10857031   0.39
[57,]     41 0.014839349 0.11155432   0.39
[58,]     54 0.013651434 0.10642943   0.39
[59,]     29 0.015185623 0.10039512   0.40
[60,]     40 0.011555232 0.11027759   0.41
[61,]     59 0.012136355 0.10563800   0.41
[62,]     86 0.012603446 0.09706857   0.41
[63,]     10 0.006637595 0.09216330   0.42
[64,]     15 0.006756635 0.10001104   0.42
[65,]     24 0.009352666 0.08889608   0.42
[66,]     63 0.011620095 0.10545954   0.42
[67,]     92 0.009113826 0.10590305   0.42
[68,]      9 0.009562369 0.09503312   0.43
[69,]     34 0.004751173 0.08109825   0.43
[70,]     35 0.006746585 0.09244119   0.43
[71,]     38 0.007042940 0.09899599   0.43
```

```
 [72,]     66  0.007429384 0.11069653   0.43
 [73,]     75  0.006104905 0.09021009   0.43
 [74,]     80  0.011267729 0.11507412   0.43
 [75,]     87  0.009361853 0.10977717   0.43
 [76,]     94  0.007624651 0.08691733   0.43
 [77,]     36  0.006841070 0.11428183   0.44
 [78,]     55  0.005923619 0.09274345   0.44
 [79,]     62  0.014148286 0.11449070   0.44
 [80,]     79  0.007690351 0.11159139   0.44
 [81,]     83  0.010178185 0.12762664   0.44
 [82,]     85  0.007351049 0.10562826   0.44
 [83,]     14  0.010226532 0.08956174   0.45
 [84,]     18  0.005114723 0.09020975   0.45
 [85,]     22  0.004356164 0.09680833   0.45
 [86,]     37  0.005390305 0.10387163   0.45
 [87,]     53  0.004877342 0.08556009   0.45
 [88,]     64  0.005560594 0.09291902   0.45
 [89,]     65 -0.001360197 0.07354842   0.45
 [90,]     70  0.005359652 0.12492498   0.45
 [91,]     96  0.007453407 0.11597641   0.45
 [92,]     19  0.006296191 0.09981632   0.46
 [93,]     61  0.006001723 0.08782904   0.46
 [94,]     95  0.003639464 0.10109827   0.46
 [95,]     99  0.006537281 0.12848510   0.46
 [96,]      5  0.006170813 0.09988789   0.47
 [97,]      3  0.007469219 0.10134040   0.48
 [98,]     20  0.011382363 0.09772801   0.48
 [99,]     98  0.005483088 0.10061177   0.49
[100,]     50  0.003712558 0.09486852   0.51
```

3.3 DBHT

```
> whas <- get.whas100.dataset()
> outliers_dbht <- survBootOutliers(
+     surv.object=Surv(time = whas$times,event = whas$status ),
+     covariate.data = whas[,2:5],
+     sod.method = "dbht",
+     B = 100,
+     B.N = 100,
+     parallel.param = SnowParam() )
> print(outliers_dbht)
```

```
      obs_id       pvalue
  [1,]     90 4.588449e-06
  [2,]     56 1.421777e-05
  [3,]     93 6.825439e-04
```

```
 [4,]     51 1.490112e-03
 [5,]      1 2.520077e-03
 [6,]     31 7.589051e-03
 [7,]      7 1.403695e-02
 [8,]      8 1.547028e-02
 [9,]     25 1.678389e-02
[10,]     94 1.760670e-02
[11,]     78 2.747890e-02
[12,]     67 3.483776e-02
[13,]     23 4.784905e-02
[14,]     45 4.901941e-02
[15,]     52 6.301141e-02
[16,]     57 7.493437e-02
[17,]     24 7.701909e-02
[18,]     88 8.310518e-02
[19,]     97 9.042274e-02
[20,]     30 1.013334e-01
[21,]     76 1.057028e-01
[22,]      6 1.185563e-01
[23,]     32 1.188880e-01
[24,]     91 1.453219e-01
[25,]      3 1.464419e-01
[26,]      5 1.587810e-01
[27,]     12 1.703916e-01
[28,]     33 1.777220e-01
[29,]     17 1.863447e-01
[30,]     46 1.939945e-01
[31,]     13 2.002493e-01
[32,]     27 2.117483e-01
[33,]     69 2.189360e-01
[34,]     29 2.237654e-01
[35,]     81 2.345347e-01
[36,]     26 2.476872e-01
[37,]     75 2.643296e-01
[38,]    100 2.829593e-01
[39,]     34 3.087835e-01
[40,]     86 3.240124e-01
[41,]     72 3.252046e-01
[42,]     22 3.335913e-01
[43,]     83 3.612813e-01
[44,]     85 3.655063e-01
[45,]     80 3.800334e-01
[46,]     82 4.095256e-01
[47,]     47 4.631899e-01
[48,]     39 4.673774e-01
[49,]     84 5.201499e-01
```

```
[50,]      68 5.356655e-01
[51,]      36 5.398887e-01
[52,]      71 5.477028e-01
[53,]      48 5.668453e-01
[54,]      62 5.680435e-01
[55,]      44 5.912894e-01
[56,]      73 6.075699e-01
[57,]      96 6.123924e-01
[58,]      64 6.512776e-01
[59,]      41 6.574869e-01
[60,]      65 6.678477e-01
[61,]       9 6.763730e-01
[62,]       2 6.899689e-01
[63,]      28 7.235870e-01
[64,]      58 7.450333e-01
[65,]      54 7.564144e-01
[66,]      11 7.731725e-01
[67,]      35 8.130895e-01
[68,]      55 8.157266e-01
[69,]      40 8.202909e-01
[70,]      89 8.241471e-01
[71,]      18 8.480330e-01
[72,]      77 8.588284e-01
[73,]       4 8.641924e-01
[74,]      61 8.739938e-01
[75,]      74 8.761325e-01
[76,]      95 8.872940e-01
[77,]      70 8.895588e-01
[78,]      37 9.147407e-01
[79,]      49 9.347676e-01
[80,]      10 9.371544e-01
[81,]      50 9.383836e-01
[82,]      63 9.492782e-01
[83,]      98 9.550096e-01
[84,]      19 9.599776e-01
[85,]      15 9.670122e-01
[86,]      99 9.722237e-01
[87,]      92 9.723669e-01
[88,]      79 9.744100e-01
[89,]      60 9.767495e-01
[90,]      14 9.844084e-01
[91,]      16 9.893585e-01
[92,]      20 9.904976e-01
[93,]      42 9.931442e-01
[94,]      53 9.933940e-01
[95,]      59 9.938306e-01
```

```
 [96,]     21 9.939427e-01
 [97,]     87 9.955295e-01
 [98,]     66 9.968771e-01
 [99,]     38 9.998473e-01
[100,]     43 9.998578e-01
```