Stats 506: PS5 Data.Table

Diana Liang

12/4/2019

The data.table package in R follows a similar pattern of data manipulation to dplyr without creating unnecessary copies of the dataset and can combine what would have been multiple lines of code into a singular line. Thus it is both more efficient in memory and in code. Below are two examples of analyses that are possible using data.tables.

Part 1: Internet in Urban vs Rural Homes by Division

This analysis, that was done before using Stata, is reproduced below using data.table. Once again Mountain South had the greatest disparity in internet access between urban and rural areas.

Disparity of Homes with Internet between Urban and Rural Areas

Division	Urban(%)	Rural(%)	Diff(%)
Mountain South	85.27 (81.32, 89.21)	66.75 (58.26, 75.24)	18.52 (7.19, 29.84)
East South Central	78.36 (70.54, 86.18)	69.03 (63.50, 74.55)	9.33 (-1.40, 20.06)
West North Central	88.00 (84.63, 91.38)	80.33 (71.49, 89.16)	7.68 (-2.45, 17.81)
Mountain North	87.42 (81.99, 92.85)	81.93 (73.82, 90.03)	5.50 (-6.19, 17.19)
West South Central	81.61 (76.41, 86.80)	76.50 (72.12, 80.88)	5.10 (-2.30, 12.50)
Pacific	88.71 (86.17, 91.25)	85.28 (77.44, 93.12)	3.43 (-4.51, 11.37)
South Atlantic	85.30 (82.63, 87.96)	82.04 (76.28, 87.80)	3.26 (-3.54, 10.05)
New England	87.57 (82.50, 92.64)	85.79 (82.36, 89.22)	1.78 (-2.48, 6.04)
East North Central	86.25 (83.76, 88.74)	86.21 (81.64, 90.78)	0.04 (-5.30, 5.39)
Middle Atlantic	89.34 (83.85, 94.82)	91.29 (85.31, 97.26)	-1.95 (-9.09, 5.19)

Part 2: Crohn’s or No Crohn’s

This example uses DNA methylation data from multiple probes placed on each chromosome to detect Crohn’s disease. The analysis below is to decide if these probes are significant in this detection. The data itself required manipulation to get rid of the 68 lines of header information about the study itself before the actual data, as well as the ending line of text.

First, proportion of probes that were significant by t-statistic were calculated by each probe group.

Proportion of Significant Probes per Probe Group

Probe Group	Proportion
ch.1.	0.0358566
ch.10	0.0310559
ch.11	0.0330579
ch.12	0.0507246
ch.13	0.0291262
ch.14	0.0930233
ch.15	0.0192308
ch.16	0.0361446
ch.17	0.0404040
ch.18	0.0322581
ch.19	0.0434783
ch.2.	0.0218182
ch.20	0.0140845
ch.21	0.0285714
ch.22	0.0434783
ch.3.	0.0204082
ch.4.	0.0222222
ch.5.	0.0416667
ch.6.	0.0333333
ch.7.	0.0312500
ch.8.	0.0671141
ch.9.	0.0363636
ch.X.	0.0333333

As shown above, ch.14 has much higher proportion of significant probes compared to all the other probes.

Next the p-values were calculated for each probe group by two-tailed, upper, and then lower significance. A significance level of 0.05 was consistently used for all three.

P-values for Two-tailed Significance

Probe Group	P-values

ch.1.	0.4655345
ch.10	0.4535465
ch.11	0.5014985
ch.12	0.2277722
ch.13	0.6213786
ch.14	0.0789211
ch.15	0.6683317
ch.16	0.4175824
ch.17	0.4445554
ch.18	0.6053946
ch.19	0.4185814
ch.2.	0.7042957
ch.20	0.7962038
ch.21	0.5424575
ch.22	0.4165834
ch.3.	0.6773227
ch.4.	0.5094905
ch.5.	0.3786214
ch.6.	0.4935065
ch.7.	0.5844156
ch.8.	0.1658342
ch.9.	0.5404595
ch.X.	0.5024975

The two-tailed p-values seem to show no probe groups were significant at or below 0.05. ch.14 is the only one that comes even close to the significance level, which is the same probe group that had an abnormally high proportion of significant probes in the previous table.

P-values for Upper Significance

Probe Group	P-values
ch.1.	0.9200799
ch.10	0.7542458
ch.11	0.8331668
ch.12	0.7372627
ch.13	0.9120879
ch.14	0.4695305
ch.15	0.8871129
ch.16	0.8781219
ch.17	0.7642358
ch.18	0.5194805

ch.19	0.5574426
ch.2.	0.6863137
ch.20	0.6463536
ch.21	1.0000000
ch.22	0.7832168
ch.3.	1.0000000
ch.4.	0.4135864
ch.5.	0.9130869
ch.6.	0.6763237
ch.7.	0.9300699
ch.8.	0.5774226
ch.9.	0.8611389
ch.X.	1.0000000

P-values for Lower Significance

Probe Group	P-values
ch.1.	0.8631369
ch.10	0.5404595
ch.11	0.6953047
ch.12	0.4805195
ch.13	0.8121878
ch.14	0.3056943
ch.15	0.6813187
ch.16	0.7092907
ch.17	0.5614386
ch.18	0.3566434
ch.19	0.3626374
ch.2.	0.4445554
ch.20	0.4225774
ch.21	1.0000000
ch.22	0.5054945
ch.3.	1.0000000
ch.4.	0.1958042
ch.5.	0.8341658
ch.6.	0.3586414
ch.7.	0.7952048
ch.8.	0.3396603
ch.9.	0.7772228
ch.X.	1.0000000

And as expected, none of the probe groups had p-values that were more extreme for either the upper or lower significant levels.

The three different p-value tables were created using different methods of computing, with the time it took to run the 1000 permutations shown below.

The first was run in a regular loop and will be used as a benchmark for the other two.

```
##    user  system elapsed 
##  155.14   87.68  298.19
```

The second was run with parallel computing using mclapply. Unfortunately Windows computers do not have the ability to use multiple cores in the way that R would like, so the timing is similar to the first run.

```
##    user  system elapsed 
##  156.60   89.30  292.36
```

The third was run with parallel computing using future to create multiple R sessions. Two sessions were used, so it makes sense that the elapsed time was cut in half.

```
##    user  system elapsed 
##    5.06    1.92  135.71
```

Overall parallel computing allows R code to run dramatically faster. This can be useful if code is expected to take long periods of time to run, since parallel computing may cut the expected time by a factor of cores and session available.