Tancredi simulation Notes

9/14/2012

|  |  |
| --- | --- |
| File Name | Purpose |
| dt\_simFunctions\_20120820.r | Contains functions used for created the simulated data, getting p-values from several types of analytic methods (getPs), run the getPs function on the various data sets (getResults), break up the error reporting from the p-values (processResults), and summarize the p-value results by getting the empirical error rates (sumzP), figure out which glymer instances have warnings (checkWarningsInternal). |
| dt\_createData\_20120820.r | Calls simFunctions. Data parameters are supplied and simulated data lists are created and saved. |
| dt\_evalData\_20120820 | Loads individual data lists, runs getResults, processes output and saves. |
| dt\_exportDataForSAS\_20120904\_server | Turns the simulated data lists into dataframes with an ID variable so they can be run in SAS |
| dt\_summarizeP\_20120820 | Runs sumzP to get error rates for each set. |
| dt\_simGlimmixMacro\_server.sas | Contains two macros, one to split up the data by ratio and one to run glimmix. The glimmix macro first runs with random effects for both individual and cluster. If there are errors, it tries with just cluster (unless that already failed), and then tries without any random effects. |
| dt\_runMacro\_server.sas | Reads in a data set, splits it by ratio, and runs the glimmix macro. It does this for all 9 data sets. |

**Data and Results (10/18/2012)**

* Data made by dt\_createData is in list form and named in the following form: listKxNySz.rdata. The big K refers to the number of clusters, N refers to the sample size per set of parameters and is always 13,000 (13k), S refers to the sample configuration. S100 means 100 observations with pre and post, 60 with just pre. S130 means 130 observations with both pre and post. S1003030 means 100 observations with pre and post, 30 with just pre, 30 with just post.
* The list is length 4. Inside each primary list (list[[ j ]]) is another list of length 13000. The sampling ratio for list[[ j ]] is the same for all simulated data sets it contains. The ratio order is 0.5, 1, 2, and 4, but you can see this by looking at the data also.
* Using dt\_exportDataForSAS, you can convert these data lists into a .csv file to be used in SAS.
* The R results are stored in several phases:
  + Temp\_resultKxN13kSz.rdata – result of getResults(), list of length 4 (one for each ratio), the internal list is of length 13,001 - the last observation indicates the ratio and sample type.
  + resultKxN13kSz.rdata – result of processResults(), list of length 4 (one for each ratio). The first element inside the primary list is a string with the ratio/sample type. The second element is a table of all the p-values for each run. The third element is the list of errors from each run.
  + pProf\_KxN13kSz.rdata – result of sumzP(), list of length 4 (one for each ratio). The first element is a matrix showing the proportion of p-values < 0.05 for each test and the number of errors encountered during the calculation. The proportion of p-values is based on the total number of cases WITHOUT errors. The second element is a list of differences between rao p-values for one-level tests and two-level tests. The third element is the difference between wald p-values, and the fourth is the difference between survey adjusted p-values.
* The SAS results just come from proc glimmix and are broken up into different files by K, S, and ratio and called something that looks like sasOut\_KxSzRq.csv. The files end in R0\_5, R1, R2, and R4 which imply ratios of 0.5, 1, 2, and 4.