```
/*
##########################################################################
##########################################################################
#<a rel="license" href="http://creativecommons.org/licenses/by-nc-
nd/4.0/"><img alt="Creative Commons Licence" style="border-width:0"
src="https://i.creativecommons.org/l/by-nc-nd/4.0/88x31.png" /></a><br
/>This work is licensed under a <a rel="license"
href="http://creativecommons.org/licenses/by-nc-nd/4.0/">Creative Commons
Attribution-NonCommercial-NoDerivatives 4.0 International License</a>.
##########################################################################
##########################################################################
#Course Authored By:
#-------------------------------------------------------------------------
----------------------------------
#K.Srinivasan
#NeuronRain Documentation and Licensing: http://neuronrain-
documentation.readthedocs.io/en/latest/
#Personal website(research): https://sites.google.com/site/kuja27/
#-------------------------------------------------------------------------
----------------------------------
##########################################################################
##########################################################################
*/

-------------------------------------------------------------------------
----------------------------------
823. (THEORY and FEATURE) People Analytics - Social Network and Urban
Sprawl Analytics - Ricker - Population Dynamics - Functions, Loops,
Variables, Timeseries and Plots in R - 22,23,24,25,26 May 2020, 2,3,4
June 2020 - related to 487,572,770
-------------------------------------------------------------------------
----------------------------------
R along with SAS is a standard BigData analytics programming language
having MATLAB capabilities. Code example Ricker.R in code/ defines a
function ricker() which computes the Ricker Chaotic Biological logistic
function for population dynamics (Nt * exp(r(1 - Nt/K)) a variant of
[Robert May] logistic for pandemic non-linear dynamics modelling (Initial
condition is per generation and K=1). The code example demonstrates
assignment of variables by "<-" operator, for loop block, list
initialization by ":", initialization of timeseries by ts and plotting
graphics by plot(). Function block by "<- function()" has been commented
which can be run as main. R scripts could be executed either by RStudio
or Rscript CLI. R Graphics plot from RStudio and logs from Rscript have
been committed.

References:
----------
1.R for Beginners - [Emmanuel Paradis] - https://cran.r-
project.org/doc/contrib/Paradis-rdebuts_en.pdf
2.Biological Logistic - [Robert May] -
http://abel.harvard.edu/archive/118r_spring_05/docs/may.pdf - equation 4

-------------------------------------------------------------------------
----------------------------------
824. (FEATURE) People Analytics - Social Network Analysis -
CoronaVirus2019 analyzer - Statistics in R - Data
Frames,Factors,Histograms,Fitting a Probability Distribution - 27 May
2020, 2,3,4 June 2020 - related to 487,572,770
-------------------------------------------------------------------------
----------------------------------
```

R is the prominent choice for data science because of immense builtins for statistical analysis. Contrived Code example FileIOVectStats.R demonstrates analysis of CoronaVirus2019 dataset from https://www.worldometers.info/coronavirus/ by computing following metrics from it:

- Histogram
- ECDF - Cumulative Density Function
- Stem
- Rug
- Find the correlation to Bell curve - Shapiro-Wilk test
- Find the correlation to Bell curve - Kolmogorov-Smirnov (KS) test
- T-two sample test
- Mean,Median,Quantiles summary

COVID2019 data requires preprocessing of the text data and R script executes the following:

- attaching a dataframe by attach()
- instantiate Data Frame after reading the data file by read.table()
- summary() from dataframe
- accessing each field of read.table() dataframe by [[]] operator
- converting string fields to numeric by as.numeric()
- splitting sample into two for T-test
- instantiate a factor object (which categorizes the data)

Statistical tests - Shapiro,KS,T-test - try to find the proximity of the dataset to normal distribution.
Histogram bucketization of COVID2019 shows an initial huge bucket followed by almost equal sized buckets.

---

825. (THEORY and FEATURE) People Analytics - Social Network Analysis - CoronaVirus2019 analyzer - Correlation coefficients in R - Concatenation of Vectors, c() function, Vector arithmetic, Correlation of two datasets - related to 487,572,770 - 2,3,4 June 2020

R provides various facilities for creating sequences - seq(), array(), c(), data.frame(), vector(). Code example VerhulstPearlReed.R analyzes the correlation between a Chaotic sequence from Verhulst-Pearl-Reed logistic law (and biological logistic) and CoronaVirus2019 dataset, both of same length. Spread of memes,fads,epidemics and diffusion of concepts in community is a Chaotic process which can be modelled by Erdos-Renyi random graphs and Cellular automata. Though Chaotic fractal datasets (Mandelbrot) are undecidable, from social network analysis literature it is evident that emergence of giant components in scalefree (Pareto 80-20 rule - number of vertices of degree r is propertional to 1/r^k - high degree vertices are least and low degree vertices are numerous) random graphs are inevitable. Approximate correlation to onset of Chaos could be deciphered by varying initial condition and bifurcation parameter(lambda) of logistic and finding maximum correlation to CoronaVirus2019 dataset. COVID2019 dataset is scaled to decimals in interval [0,1] by vector division feature of R which applies division by maximum element to each element of sequence. Vector division could also be performed on dataframe objects (commented). Scale-normalized dataset is then correlated to Chaotic sequence by Kendall,Spearman,Pearson coefficients (~45% maximum). Concatenation is done by c() function within for loop which reassigns the vector after concatenating next element. Double precision arithmetic has been coerced. Factor objects of both logistics are printed which show the levels of data. Graphics are plotted by timeseries plot() function (most recent to the left) - Chaotic sequence oscillates heavily by period

doubling. Low p-value implies null hypothesis (no correlation) can be rejected with high confidence. Maximum correlation occurs at lambda=4.0 and initial condition=0.00000000000001 and for increasing lambda both logistics coincide (resemble a heaviside step function). Return value is a concatenated vector by c().

References:

825.1 R and S documentation - https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/c

---
828. (THEORY and FEATURE) People Analytics - Social Network Analysis - CoronaVirus2019 analyzer - Linear Models - 11,12 June 2020 - related to 487,572,752,770,823,824,825

R has variety of builtins for Linear Regression and Logistic Regression - Linear Models - lm() and glm() functions. Code example LinearModels.R computes linear model lm() and generalized linear model glm() for CoronaVirus2019 dataset which correlate per-day and total fields by various family of fit measures - gaussian, gamma, poisson, binomial and link functions - logit, identity, log, inverse, 1/mu^2. It also demonstrates matrix creation by cbind() which combines vectors and $ notation for dataframe fields. Summary including linear.predictor for each model is printed to testlogs/LinearModels.log.11June2020. Pandemics as Social network fad diffusion are traditionally fit to Chaos, Cellular automata, ERSIR random graph model, SIS random graph model, exponential or poisson distributions. An example Theoretical regression model which explains the spread based on population density (average degree in urban sprawl social networks) could be:

Per day Spread = weight1 * total_infected - weight2 * total_recovered - weight3 * deaths + weight4 * recovery_time ... + constant

for variables total_infected, deaths, recovery time and total_recovered and weights weights3=1, weight4, weight1=average_degree and weight2=1 (spread is en masse and depends on degree of network and recovery time but recovery benefits only individual). By CAGraph social network concept diffusion model, weight1=average_degree=8 for 2 dimensional cellular automaton increment growth rule. Linear model is dependent mostly on recovery time which then becomes:

Per day Spread = 8 * total_infected - 1 * total_recovered - 1 * deaths + weight4 * recovery_time ... + constant

---
831. (THEORY and FEATURE) People Analytics - Social Network Analysis - CoronaVirus2019 analyzer - Cellular Automaton Graph (CAGraph) - Regression Models of Diffusion - Iteratively Re-Weighted Least Squares (IWLS) - 22,23,24,25 June 2020 - related to 678,740,762,828,830 and all sections on Business Intelligence, Voting Analytics, Urban Sprawl Analytics, Random Walks on Expanders of NeuronRain Theory Drafts
---
1.In continuation of R CoronaVirus2019 models earlier, Cellular Automaton Graph diffusion logit in 828 has been implemented by glm() and lm() utilities of R library.

2.Two R functions in CAGraphLogit.R - cagraphlogit() and cagraphprojections() - respectively learn a regression model from COVID19 data and project it to a later datapoint to obtain per day global spread.
3.glm() and lm() functions learn the model:

```
Perdiem ~ Infected + Deaths + Recovered + RecoveryTime
```

based on Susceptible-Infected-Recovered (SIR) data gathered from few giant geographic clusters.
4.Logs of code/testlogs/CAGraphLogit.log.24June2020 detail the predictors, coefficients and fitted model.
5.Regression model coefficients are learnt by Iteratively Re-Weighted Least Squares implementations of glm() and lm(). CoronaVirus2019 dataset has been chosen as a representational dataset because of its unprecedented global scale and randomness which could simulate any other random process involving people including business, economy and majority voting (Preferential attachment of high market share brands versus Double Jeopardy of low market share brands - new products gain market share by word of mouth and promos, spread of opinions as opposed to pandemics decide elections than individual rational decision making - correlated majority or statistical dependence of voters - CJT for Correlated Majority - https://www.sciencedirect.com/science/article/abs/pii/016726819400068P , Markets' spikes and corrections are explained by Bounded rationality and Irrational exuberance).
6.All learnt weights have been multiplied by 400 except total_infected which is multiplied by 8*400 = 3200 as 2-dimensional Cellular Automaton Graph heuristics for High degree Expander graphs (from Section 740, degree 400 graphs are Expanders which facilitate huge spread subject to |eigenvalue(k)|/d-regularity <= 1/10) - Rationale being Spread in an Urban Sprawl social network is a Random walk on Expander Cellular Automaton Graph. Function cagraphprojections() computes the following global spread per day from the learnt model applying signs (and multiplications by 8 and 400) for each dependent variable of recent COVID19 SIR data (by abs() of weights to ignore signs and without abs()). Model could be refined by incorporating additional dependent variables:

```
[1] "Per day spread(abs):"
[1] 121087.5
[1] "Per day spread:"
[1] 149975.7
```

7. Section 830 has a derivation for number of possible arrangements or 2-colorings of people subject to avoidance criteria. Number of possible arrangements of people (uniquely identified by integers) bijectively equal number of possible complementary equations or complement diophantines defined on the integer sequences (Complementary Equations, Functional equations, Beatty sequences, Taichi - partition of Integer sequences - https://cs.uwaterloo.ca/journals/JIS/VOL10/Kimberling/kimberling26.pdf). In the context of diffusion in social network, 2-coloring reduces to Infected-Uninfected. An example combinatorial avoidance: Infected and Uninfected couldn't be juxtaposed.
8. Section 762 defines least square model for variables in business and econometric datasets - By replacing the variables of COVID19 analyzer, similar models could be learnt for those datasets as well.

---

870. (THEORY and FEATURE) People Analytics - Social Network Analysis - CoronaVirus2019 Analyzer - Exponential Fit in R - related to 744,831 and all sections on Chaos, Cybercrime analyzers, Game theoretic Cybersecurity, Pseudorandomness, Random walks on Cellular Automata and Expander Graphs - 11 October 2020, 12 October 2020

---
1. CoronaVirus2019 dataset has been thus far fit to non-exponential probability distributions and Chaotic non-linear models.
2. Chaotic models suffer from theoretical limitations of computational undecidability thereby prohibiting accurately learnt models.
3. This commit implements exponential fit of CoronaVirus2019 dataset till 11 October 2020, by lm() applying logarithmic version of:

$$Fatalities = A*exp(B*days)$$

which is log2(Fatalities) = logA + B*days (R linear model is used for learning non-linear exponential model)
4. Logs code/testlogs/Exponential.log.12October2020 print the exponential fit model.
5. Previous exponential model is the solution for differential equation d(Fatalities)/Fatalities = B*d(days) which is similar to DEs in Erdos-Renyi Susceptible-Infected-Recovered random graph epidemic model.
6. Differential equation in (5) implies rate of change of fatalities depends on instantaneous fatalities (Geometric progression or evolving m-ary tree).
7. From coefficients in logs, COVID19 dataset is fit to A=11.27683950 and B=0.04250611