Article can be accessed online at http://www.publishingindia.com

Data Mining and Analysis by R Language for Business Research: A Case Study on Stress and its Influence on Health

Kamakshaiah Musunuru*

Abstract

R is not only a statistical suite but also efficient data mining software for data manipulation, calculation and graphical display. In fact, R being a language also has an effective data handling and storage facility. Besides having a suite of operators for calculations on arrays, in particular matrices. The R is developed from a simple and effective programming language (called "S") which includes conditionals, loops; user defined recursive functions and input and output facilities. Methods: In this paper the data mining capabilities of R has been explained with the help of a study on secondary data sources, obtained from certain authenticated sources. The study is all about to understand stress with respect to certain other factors like heavy drinking, perceived health and life satisfaction. As it mentioned the data so used is secondary in nature, which is in its crude from having no sense to the user. But by a systematic execution of certain data mining tools, like correlation and MANOVA, certain important relationships along with ties were realized. Conclusions: The realizations were that all variables are strictly correlated with Karl Pearson correlation coefficient ranging from 0.73 to 0.99. In significant test all variables do not belie with alternative hypothesis, which means the association/ relationship is not zero. In MANOVA, the null hypothesis is rejected as the p-value is less than 0.05. Apart from this, most interestingly the variables are behaving like cohorts whereby resulting cohort effect.

Keywords: R Language, Rstudio, Secondary Data, Data Mining, Correlation, Manova

What is R?

R is an integrated suite of software that handles calculation, data manipulation and graphical display. (Venables and Smith, 1999) some of the main feature of R (perhaps, above and over others) are: (1) effective data handling and storage facility, (2) a large, coherent, integrated collection of intermediate tools for data analysis, (3) a suite of operators for calculations on arrays, in particular matrices, (4) a matured and suitable programming language based on "S", (5) graphical facilities for data analysis and display by either in computer or on hard-copy in the form of reproducible research and etc. (Venables and Smith, 1999) In fact, R is only an implementation of S which was developed in Bell Laboratories (of AT&T, now Lucent Technologies) by Rick Becker, John Chambers and Alan Wilks, way back in 1970's for UNIX operating system, the same individuals also forms as basis for S-Plus statistical system.[1] Indeed, R project was initiated by two wonderful individuals known as *Robert Gentleman* and *Ross Ihaka* (that is why it is called as R) in department of statistics, University of Auckland in 1995. (Owen, 2010) Currently, taken care by core-group (international group of volunteers and developers). The R is freely available for download from:

R and Other Statistics Suites

R is *lingua franca* of statistics. In fact, it is wise to refer it to as developing environment where statistical implementations were done. There are two different devices for implementations they are *the base* and *the packages*. The base do consist of few functions along with data packages; whereas, the packages are special software which can be obtained inside R as per developers needs. There are around 3000 packages as

[1] S-Plus is an commercial suite of S distributed by Insightful Corporation http://www.r-project.org or alternatively from (US): http://cran.us.r-project.org/

* Associate Professor, Business Analytics, Dhruva College of Management, Hyderabad, India.
Email: kamakshaiah.m@gmail.com

Table 1: Differences among few business analytical software

Name	*Advantages*	*Disadvantages*	*Open Source*	*Typical Users*
R	Library Support; Visualization, Steep learning curve	Minimal GUI	Yes	Finance and statistics
Matlab	Elegant matrix support; Visualization	Expensive and incomplete statistical support	No	Engineering
SciPy/NumPy/MatPlotlib	Python general purpose programming language	Immature	Yes	Engineering
Excel	Easy, Visual and Flexible	Large data sets	No	Business
SAS	Large data sets	Expensive, Out dated programming language	No	Business
Stata	Easy statistical analysis	-	No	Science
SPSS	Link Stata but more expensive and worse	-	-	-

Source: O"Conner, B. from http://www.thejuliagroup.com/blog/?p=1757

on today, for diverse needs of both users and developers. Many of these wonderful software packages are available from CRAN (Comprehensive R Archive Network), while others are project based.[2]

There are umpteen number of software suites for statistical analysis, each having their own advantages.[3] One of the major alternatives to R language is SPSS. Although, SPSS is ruling the roost, it is mainly due to pioneering effect not by stamina. In fact, SPSS can not be comparable with R language for they tend to differ in terms of user approach. One of the key difference is that SPSS is being proprietary in nature is not free and require to purchase license, and the user needs to shell out huge amount of money. Some of the other differences can be; few users complain about sloppy and insufficient analytical stamina of SPSS, besides poor community help. (Stackoverflow.com; Henrick, 2010) Perhaps, R is meant for programmers, which means, advanced users who might not be novices in statistics and data mining, whereas, SPSS being GUI, is recommended to beginners; due to which, the SPSS syntax seems to be terrible compared to R. (Henrick, 2010) As far as, graphical outputs are concerned, R has more options as per customized needs of the user compared to SPSS, since, SPSS is supplied with vendor-lock-in. [7] Above all, R is not for mere analysis but possess very efficient ways to reproducible research, the user easily can prepare ready readable reports with help of certain reporting software, that are absolutely free. The following illustration could sort out about the differences among few major software giant□s that are currently in use.

O"Connor, (2009), asserts that, in fact, all these software can be categorized as; (1) programming oriented solutions, and (2) non-programming oriented solutions (he calls non-programming suites as analytical). [8] He further adds that R, Matlab, Python are programming based and Excel, SAS, SPSS and etc. are analytic based, but regarding comparison, he affirms that R is very efficient in contrast to the rest. He reports athat Matlab"s language is week, but best for mathematical algorithms. While SAS is used by older crown, is also preferred over SPSS due to its cheaper trait compared to the former. Another important issue is any tool need to be user friendly in terms of needs, but not alone on maneuverability. (Friedman, n.d.).

How R is Effective in Data Mining?

The below study on certain health data sets, that are available from *statistics Canada* (described in research

[2] In fact, bioconductor (www.bioconductor.org/), psych (www.personality-project.org), have, FactoMineR (http://factominer.free.fr/), EnQuireR (http://enquirer.free.fr/) and etc. are to name a few, that they have their own project based efforts, from where users can obtain all the data regarding installation and usage, apart from being available from CRAN.

[3] For more information on SPSS, please visit: www.ibm.com/software/analytics/spss/

methods) shows us as how R can be helpful in mining data to business researchers. The following is the rational to the study.

What is stress? And how it affects health?

It appears that the real meaning of stress is still difficult to comprehend. Although, we talk about stress many times in real life, it is still a nascent area for research. Mostly, stress is considered to be antecedents of some happening which is not expected (Castro, F. G. 1989). Some think that stress is thinking about unwanted happenings in life, others interprets as a cognitive or physical reaction to an event (perhaps both). Although, mind and body play enormous roll in getting stressed, it is indeed, the thought that is most important aspect, perhaps which determines the stress. (Jacob, 2001; www.bam.gov) It is the sin, but not the sinner, might be a typical approach to understand stress.

In fact the stress is a cognitive process, which might trigger certain bodily responses to a given stimulus. (Ursin andEriksen, 2003; MathewsandMackintosh, 1998) Why somebody sweats or shivers when anticipating a danger? Why someone tends to be emotional, when something hurts his/her ego? Why someone grow anger, when he/she feels hungry? Why someone could not sleep well when denigrated by some other? These are some typical questions, answers to which might perhaps reveal abnormal nature of stress. Sometimes the situations are stressful, might not be cognitive in nature, but it might be due to lack of sufficient preparations or coping skills.

How we understand stress and stress provoking situations, perhaps impact our health. (www.cmha.ca) A careful observation upon events commonly known as stressors and individual cognitive reactions to the same might help to gain control over stress. While some typical symptoms of stress are; pressurization, anxiety, memory loss, ulcers, insomnia and etc. and effects can be mental health problems, cardiovascular disease and etc. (www.cmha.ca) Yoshitaka (2006) in his study on aboriginal tribes founded that the stress is prevalent and play significant role in every body‟s life. He also argues that stress is not only creates health related issues but emanates at different intertwined levels like socio-economic, cultural, historical, political and etc. apart from finding various reasons, he also suggests certain stress coping strategies like; collective strengths, gaining strengths through spirituality, cultivating cultural identity, using personal/individual aspects of strengths and making positive transformations in meaningful ways. Burton (2008), in his case on healthy workplace, while proposing certain avenues that makes workplace more interesting to an individual, emphasizes on *stress management training* as one of the important avenues that an employer provides to an individual to feel better. In fact, he exhorts that individual‟s mental and physical well-being depends upon identifying workplace stressors and getting them ready to face them through stress management.

There are other studies which emphasize on certain important factors of society. If the health depends on perceived stress of an individual, then where from this stress arises in individual? (Pearlin, 1989) certainly the primary cause can be society. The conceptual and methodological tools of social stress theory and research are relevant in investigating the pathways linking social structure and health (McDonough, et al.,2002). They argue that social roles and social economic positions are consequential for health because they are responsible for daily life struggles, these struggles in turns might responsible for stress.

Objectives & Hypothesis

The following are the objectives to the study:

1. To know about effectiveness of secondary data and its efficiency in deriving patterns
2. To assess the power of R in mining data such as from crude and raw set of data.
3. To study about stress and its influence on other factors like heavy drinking, health perception and life satisfaction.
4. To know and aware that whether the effect of stress could be same across various age groups and gender.

Research Methods

R is very strong as far as statistical tools are concerned; it is very demonstrative while working on visualization tools, besides, there are certain efficient add-on packages like, *diagram, rgl, dingraph, ggplot (1,2, etc.),* scatterplot3D and etc., but visualizations in this paper worked by the basic R graphic engine, viz. lattice. Apart from simple

tables and graphs (descriptive analysis), MANOVA is used in order to test the hypothesis and for further analysis. In MANOVA, the test of hypothesis is a default mechanism, apart from that there is a provision for further analysis as test of fitness, test of normality, test of significance and leverage analysis. In MANOVA the null hypothesis could be:

$$H_{0:}\ \mu_1 = \mu_2 = \mu_3 = \mu_4$$

Where μ is a vector of means for a given number of dependent variables, here, the dependent variables are *heavy drinking, perceived health and life satisfaction.* This study is carried out under the basic assumption that the above mentioned variables may tend to depend on *perceived life stress.* Hence, μ_1, μ_2, μ_3, μ_4 represents the vectors of means for a given variable respectively. In other words; this could be as below:

$$H_o = \begin{bmatrix} x_{11} \\ x_{12} \\ \cdot \\ \cdot \\ x_{p1} \end{bmatrix} = \begin{bmatrix} y_{12} \\ y_{22} \\ \cdot \\ \cdot \\ y_{p2} \end{bmatrix} = \begin{bmatrix} z_{13} \\ z_{23} \\ \cdot \\ \cdot \\ z_{p3} \end{bmatrix} = \ldots = \begin{bmatrix} A_{1n} \\ B_{2n} \\ \cdot \\ \cdot \\ Z_{pn} \end{bmatrix}$$

Where *p* represents the total number of dependent variables for *k* levels, our test statistic will be *Λ* and is computed as:

$$\Lambda = \frac{|W|}{|T|} = \frac{|W|}{|B+W|}$$

Where *W* and *T* cross product matrices of sum of squares within, *B* is between effect, if *B* is very large then *Λ* approaches zero, however, if B is very small or zero then *Λ* approaches to one.

To put it simply; $W = W_1 + W_2 + W_3 + W_4$ for all respective variables; where

$$W_1 = \begin{bmatrix} SS_1 & SS_{12} \\ SS_{21} & SS_2 \end{bmatrix}$$

And

$$SS_1 = \sum_{j=1}^{n} (y_{1j} - \bar{y}_1)$$

$$SS_2 = \sum_{j=1}^{n} (y_{2j} - \bar{y}_2)$$

$$SS_{12} = SS_{21} = \sum_{j=1}^{n} (y_{1j} - \bar{y}_1)(y_{2j} - \bar{y}_2)$$

In general worlds; the following can be hypotheses to the study:

H_0 = There is no significant difference among variables under the study

H_1 = There is significant difference among variables under the study

The research is primarily *descriptive* and *inferential* in nature.[4] The descriptive part of the study could explain who, what, when, where and how, whereas, the inferential analysis is all about scientific approach to prove or disprove the research proposition. The research attempt is to ascertain that if any logical relationships exist among stress and other factors of the study. Hence this study deals with description and exploration rather than finding cause and effect among variables.

Due to certain operative constraints the data is restricted to be secondary. Apart from its limitations, the secondary data also has certain strengths. Two of the potential advantages of secondary data are its *economy* and *time*. (McDonough, et al., 2002) Instead of using research resources in this very phase (data collection); they might perhaps be employed efficiently in other important phases of research process. The data is obtained from *statistics Canada*; a member of the *United Nations Statistical Commission* and *Industry Portfolio* of Canada.[5]

There are four variables that were taken to the study, viz., *perceived life stress*, *heavy drinking*, *perceived health* and *life satisfaction*. The data set is organized by showing variables as columns; *age* and *gender* as row-wise individuals, by keeping the whole data set as 4X15 order matrix.

R Language along with *RStudio* was used as statistical software for analysis. *R* Language is regarded as one of the efficient software's in academics and industry for statistical analysis. (Fox, 2005) At very outset the analysis is linear modeling, which further extended to be M(ANOVA) with giving priority to test of normality and

[4] The meaning to the term descriptive is given differently in different sources of text.

[5] Statistics Canada is member of the industry portfolio, which is an official website that provides data on very vital aspects in Canada. Industry portfolio is a consortium of eleven federal departments and agencies in Canada.

Table 2: Stress dataset

Age and Gender	*Perceived Life Stress*	*Heavy Drinking*	*Perceived Health*	*Life Satisfaction*
15 to 19 years	14.20279	23.14609	49.05717	124.5221
Males	5.38073	14.22287	25.58851	64.10982
Females	8.82206	8.92321	23.46866	60.41232
20 to 34 years	61.85805	98.08602	99.95336	245.5136
Males	28.70309	66.3883	50.75036	122.9862
Females	33.15496	31.69774	49.20299	122.5274
35 to 44 years	56.56169	41.49664	70.44681	175.0482
Males	27.35909	31.15958	34.64086	87.15947
Females	29.20261	10.33706	35.80596	87.88873
45 to 64 years	95.58564	65.58711	128.1213	325.2061
Males	45.2948	49.51727	62.45136	160.4888
Females	50.29082	16.06985	65.6699	164.7173
65 years and older	19.2188	9.28648	55.78105	144.0355
Males	8.05141	7.6707	25.22971	64.02301
Females	11.16738	1.61577	30.55135	80.0125
Minimum	5.381	1.616	23.47	60.41
1st Quartile	12.685	9.812	32.60	83.59
Median	28.703	23.146	49.20	122.99
Mean	32.990	31.680	53.78	135.24
3rd Quartile	47.793	45.507	64.06	162.60
Maximum	95.586	98.086	128.12	325.21

Source: Statistics Canada

leverage analysis. Since, the analysis multivariate, the following is the underlying linear model to the analysis:

Analysis and Discussion

The study was done on health data of individuals that are grouped in 5 categories in terms of age and gender. There are five age groups namely; 15 to 19 years, 20 to 34 years, 35 to 44 years, 45 to 64 years and more than 65 years respectively in the dataset. These categories are again segregated in terms of male and female. The table 2 shows the summary of dataset:

In the above table the last six rows were not the part of the data set, in fact, they have been obtained by the following code. The name of the data set is *stress*; the following code can read the concerned file and give summary statistics (the last six columns of the above table).

```
> stress = read.csv(file.choose())
> stress
> summary(stress)
```

In fact, the actual output does consist of abundant of information, but for the sake of simplicity only required data is used for explanation.

Correlation Analysis

Although, summary statistics could give umpteen amounts of insights to the data set, but due to its simplicity summary may not be sufficient to deal with the phenomenon. Hence, There further analysis of correlation perhaps could reveal more information or add sense to the data. Basically the correlation analysis is all about relationships among study variables, which means, how variables are explained by variance. The following code can execute correlation on data set.

Table 3: correlation matrix

Variable	*Perceived life stress*	*Heavy drinking*	*Perceived health*	*Life satisfaction*
Perceived life stress	1.0000000000	0.7099653785	0.9418687127	0.9391121185
Heavy drinking	0.7099653785	1.0000000000	0.7553249222	0.7372740615
Perceived health	0.9418687127	0.7553249222	1.0000000000	0.9992416240
Life satisfaction	0.9391121185	0.7372740615	0.9992416240	1.0000000000

Source: statistical analysis on dataset

Table4: correlation significant test

Variables under test	*t-value*	*p-value*	*Confidence interval*	*Estimate (r)*	*Hypothesis (true)*
Perceived Life Stress	27.3454	0.0001073	0.9684981- 0.9998748	0.9980001	Alternative
Heavy Drinking	5.0344	0.01511	0.3821661- 0.9965084	0.9456018	Alternative
Perceived Health	9.0415	0.002857	0.7481813- 0.9988735	0.9821411	Alternative
Life Satisfaction	9.321	0.002614	0.7610624- 0.9989389	0.9831698	Alternative

Source: from statistical analysis on dataset

```
> cor(stress[,2:5])
```

The relationships or associations are observed to be strictly strong and positive. The relationship in between *perceived health* and *life satisfaction* is very high as the correlation coefficient is observed to be 0.9992416240. The same sort of relationships were discovered among *perceived life stress* and *perceived health* and *perceived life stress* and *perceived life satisfaction*. The correlation coefficients are observed as 0.9418687127 and 0.9391121185, which is ironical. Does it mean that individuals who feel stress tend to feel better about their health? More inquiry is necessary to know about the relationship in between these two variables. Although the relationship in between *perceived life stress* and *heavy drinking* is fair, it is observed to be 0.709, which is the least observed among all pairs. It might be that *stress* leads to *heavy drinking* but might not be much important compared to other two variables. Hence, the model can better be explained in terms of *contributions*, i.e. variance, rather than relationships, which means, significance of individual responses to one variable might be high towards other variable. The following table gives the summary of correlation (matrix) analysis:

In case, after all, finding relationship or associations either strong or weak, if the researcher or data miner would like to find the significance of the relationship. It might be easier in *R*, the below command could give all the data necessary to draw inferences upon correlation.

```
> cor.test(stress[,2], stress[,3])
```

Under Pearson"s product movement correlation coefficient method; the alternative hypothesis for all pairs were proved true (inference). The following is the table of significance test on correlation coefficient (*r*).

Still our dilemma that the genuineness of relationship is not unraveled. Hence, more and more training on data set is required, in order to ascertain these ironic relationships. Linear models like MANOVA, perhaps, can solve this dilemma or ambiguity. One of the potential reasons for using MANOVA is being variable interaction. The following code can generate umpteen amount of information to the miner:

```
> stress.manova = manova(cbind(stress$PLS,stress$HD,
stress$PH, stress$LS)~stress$AnG, data=stress)
> summary(stress.manova)
> spirit.anova=anova(stress.manova)
> summary(spirit.anova)
> spirit.anova
> stress.aov = aov(stress.manova)
> summary(stress.aov)
```

Table5 can illustrate the output of the above R code or chunk. The F statistic of the output is 2.5721 at p-value being 0.01008 at one percent confidence (0.01). Hence, the null hypothesis is accepted.

All p-values are at 1% confidence level. All the values are

Table 4: MANOVA results on study variables

Levels	*DF*	*F Statistic (Pillai)*	*Approx. F*	*p-value*	
Age and Sex	6	2.5791	2.4201	0.01008	
Summary of MANOVA					
Response	**DF**	**Sum of Squires**	**Mean Squires**	**F value**	**P-value**
Perceived life stress	6	6561.8	1093.64	3.8925	0.04037
	8	2247.7	280.96		
Heavy drinking	6	7863.7	1310.61	3.6259	0.04844
	8	2891.7	361.46		
Perceived health	6	9773.8	1628.96	5.9418	0.01234
	8	2193.2	274.15		
Life satisfaction	6	61482	10247.0	5.9858	0.01207
	8	13695	1711.9		

Source: statistical analysis on dataset

more than 0.01, hence significance is very low. But a very close observation could bring more interesting findings that, in spite of clear indication of global p-value that the difference is significant, this significance is not so strong

with respect to first two variables i.e., *perceived life stress* and *heavy drinking*. Hence, it is conspicuous that there is further possibility of grouping of these variables, viz., *perceived life stress* and *heavy drinking* are close together in terms of their p-values; *perceived health* and *life satisfaction* might be grouped together in terms of their respective p-values. Perhaps the variables are subjected to *cohort-effect*.

Conclusions & Summary

The R language is a very effective data mining software. From the empirical study it is realized that a systematic execution of certain statistical models can add value or make sense of data such as secondary data, which is in the form of mere characters or crude symbols.

Regarding the study, the analysis on stress vs. its effects revealed that there are perfect relationships among the variables; all variables are so strong in association. The *Karl Pearson coefficient of correlation* is fair with maximum of 0.99 and minimum of 0.73. Even in significance test it was observed that the all null hypothesis were rejected (the null hypothesis is being that there is no correlation in between the variables under study). As far as the *age* and *gender* (levels) are concerned; there is no significant difference across different age levels and gender with respect to variables under study (although the difference is not so significant). More interestingly there are certain sub-groups (pattern recognition) among the variables (cohort effect). *It needs further research to explore why certain variables are behaving as cohorts*.

References

Venables, W. N., & Smith, D. M. (1999). Introduction to R:Notes on R. ISBN: 3-900051-12-7

ibid.

W. J. (2010). The R guide. Retrieved from http://www.mathcs.richmond.edu/~wowen/TheRGuide.pdf

Retrieved from http://stackoverflow.com/questions/3787231/r-and-spss-difference

Henrick. (2010). R and SPSS difference (forum discussion on difference). Retrieved from http://stackoverflow.com/questions/3787231/r-and-spss-difference

ibid.

ibid.

O"Conner, B. (2009). Comparisons of data analysis packages: R, Matlab, Scipy, Excel, SAS, SPSS, Stata. Retrieved from http://www.thejuliagroup.com/blog/?p=1757

Friedman, J. H. (n.d.). Data mining and statistics: what‟s the connection? Retrieved from http://www- stat.stanford. edu/~jhf/ftp/dm-stat.pdf

Castro, F. G., Maddahian, E., Newcomb, M.D., & Bentler, P. M. (1989). A multivariate model of the determinants of cigarette smoking among adolescents. *Journal of Health and Social Behavior*, *28*, 273-289.

Jacob, G. D. (2001). The physiology of mind-body interactions: the stress response and the relaxation response. *The Journal of Alternative and Complementary Medicine, 7*(1), s-83-s-92.

The Body-mind Connections of Stress. Retrieved from http:// www.bam.gov/teachers/activities/stress_body_mind.pdf

Ursin, H., & Eriksen, H. R. (2003). *The Cognitive Activation Theory Of Stress*. Retrieved from http://meagherlab. tamu.edu/M- Meagher/%20Health%20Psyc%20630/ Readings%20630/Stress%20readings/cog%20stress%20 04.pdf

Mathews, A., & Mackintosh, B. (1998). A cognitive model of selective processing in anxiety. *Cognitive Therapy and Research., 22*(6), 539-560.

Coping With Stress, Canadian mental association, Retrieved from http://www.cmha.ca/data/1/ rec_docs/403_CMHA_coping_with_stress_EN.pdf

ibid

Yoshitaka, I. (2006). Stress coping among aboriginal individuals with Diabetes in an Urban Candian City: From Woundedness to Resilience. *Journal of Aboriginal Health*, *15*

Burton, J. (2008). The business case for healthy workplace. *Industrial Accident Prevention Association.* Retrieved from http://www.iapa.ca/pdf/fd_business_ case_healthy_workplace.pdf

Pearlin, L. I. (1989). The sociological study of stress. *Journal Of Health And Social Behavior, 30,* 241-256.

McDonough, P. Walters, V., & Strohschein, L. (2002). Chronic stress and the social patterning of women‟s health in Canada. *Social science and medicine*, *54* (spl.), 767-782

McDonough, P. Walters, V., & Strohschein, L. (2002). Chronic stress and the social patterning of women‟s health in Canada. *Social Science and Medicine*, *54* (spl.), 767-782

Fox, J. (2005). The R commander: a basic statistics graphical user interface to R. *Journal of Statistical Software, 14*(9), 1--42.