Western University
Scholarship@Western

Undergraduate Honors Theses

Psychology

Winter 4-30-2014

Exploring big data analytics: Predictors of psychological well-being in a Canadian sample

Canaan Legault
King's University College, clegaul5@uwo.ca

Exploring Big Data Analytics:

Predictors of Psychological Well-Being in a Canadian Sample

Canaan Legault

Honours Thesis

Department of Psychology

King's University College at *The* University *of* Western Ontario

London, Canada

April 2014

Thesis Advisor: Imants Baruss, BSc MSc PhD

Abstract

The current study examined the emerging field of big data analytics to assess the utility of this technology for psychological research. A study looking at predictors of psychological well-being within Statistics Canada data sets was designed as a means of testing the capabilities of standard statistical software (SPSS) for a large data set (N = 681,578). Results from this study found health, stress, relationships, income, diet, exercise, education, and alcohol abuse to be predictors of psychological well-being. SPSS was demonstrated to be able to analyze large data sets, however, actual big data analysis (terabytes to petabytes) may be unreasonable due to performance issues. Possible deterrents that could explain why big data research is not being embraced by psychologists were identified through an exploratory look at integrating these technologies into psychological research. Shortcomings of current big data technologies in relation to psychologists are discussed as well as directions for future research into big data and psychological well-being.

Exploring Big Data Analytics:

Predictors of Psychological Well-Being in a Canadian Sample

Advancements in technology are constantly yielding new areas of research for psychology, as well as new tools which can be used to discover novel insights and expand the scope of research. Big data, along with big data analytics tools, provide just such an opportunity to the field of psychology should this resource be properly utilized. There is no standard definition for what constitutes big data but the term is generally applied to sets of data which, due to their size, cannot reasonably be managed by usual data management or analytical tools (Tien, 2013). Ward (2013) explained that many of the statistical programs typically utilized by researchers, such as Statistical Package for the Social Sciences (SPSS) or Stata, are not able to process large data sets in a timely manner and would therefore limit what data a researcher could examine. It is because of this problem that new technologies and software, such as big data analytics programs like Hadoop, have emerged to overcome the limitations of traditional programs (Deroos, Deutsch, Eaton, Lapis, & Zikopoulos, 2012). Being able to utilize tools to analyze big data and work with it opens up new possibilities in regards to research and is something that, at this point in time, is not being fully realized in psychology or the other social sciences (Ovadia, 2013) but has great potential for future projects. Therefore, it is necessary that more awareness be given to this topic in psychology and further research should be done to demonstrate how the field of psychology could use these tools and what could be done to promote more social science involvement in this area.

What is "Big Data"?

The difference between big data and traditional data is typically characterized through the dimensions of volume, variety, and velocity (Deroos et al., 2012). Volume refers to the size of big data, which is generally in the range of terabytes or larger, though with technological advances the size of big data is constantly expanding (Tien, 2013). Variety refers to how big data comes in several different formats with distinctions being made between structured data, which is organized, and unstructured data, which may be messy or unorganized (Deroos et al., 2012). Velocity refers to the manner and speed at which new data can be received for analysis, such as through the Internet. Big data, according to these criteria, refers to data that is large, diverse, and growing. However, it should be mentioned that sets of big data will likely always vary in regards to what extent the concepts of volume, variety, or velocity are present. As was previously stated, there are no formal rules for what constitutes big data (Tien, 2013) so the explanation given by Deroos et al. (2012) represents a common conceptualization of what big data might look like.

Ethical Concerns for Large Data Sets

For academics, having the ability to access vast quantities of data with relative ease is a boon for research. However, this increase in accessibility, size, and storage also brings with it concerns regarding privacy and ethics. As the amount of data associated with a single participant grows it becomes easier for an individual to ascertain the participant's identity (Tien, 2013). Schadt (2012) said that because of the increasing amounts of data collected and stored there is growing concern over anonymity in the context of research in the digital age. DNA and genomic databases, GPS information, and social networking data all pose risks in regards to the identification of participants. It is paramount that big data projects continue to preserve the anonymity of participants and be mindful of the potential risks associated with larger data sets.

Not only is anonymity a concern for big data, but so is the issue of consent. This is especially true for social networking sites or services which aggregate user accounts and information from a variety of sources (Oboler, Welsh, & Cruz, 2012). Data that may be freely available, such as from Twitter or Facebook, is of growing ethical concern to researchers in regards to what “public information” can be used and how. The most prominent issues involving big data and privacy concerns stem from governmental or private corporate entities (Lesk, 2013), but it is beneficial for researchers to consider the ethical issues related to these data sets. Oboler, Welsh, and Cruz (2012) remarked that social networking sites and other repositories of large scale data offer researchers valuable data pools, but care must be taken to protect sensitive information and thought given to what consent is required. These issues can be mitigated by using databases which collect data for the purposes of research and have followed proper ethical guidelines. Having a data set where steps have been taken to ensure anonymity, consent (such as an opt-in program), and protection with respect to online security will help reduce the ethical concerns that may be associated with big data or similar future research.

Uses and Applications for Psychology

Big data has largely not been utilized within the social sciences to the same extent that it has in biomedical research (Miller, 2012) and corporate or security applications (Lesk, 2013). Aside from some industrial and organizational uses (Davenport, Barth, & Bean, 2012), social networking projects (Patterns and Predictions, 2014), and forensic applications (Collins, 2013), big data is relatively obscure in the field of psychology. The psychological applications of big data and big data analytics are somewhat limited in the current literature, with many projects in the pilot stages or research still currently ongoing. The area where big data has been most incorporated is in industry, where corporations use it to better understand business environments, consumer behaviour, and customer concerns or demands (Davenport, Barth, & Bean, 2012). Big data can aid in various levels of the corporate process that psychology is concerned with. Aspects such as marketing, advertising, and decision making can all benefit from the insights contained within large data sets. Corporations often have a dearth of information at their disposal, but the problem is they are not effectively using it to their advantage through the proper analysis (Koh, 2012). This technology has great potential for businesses that chose to utilize big data analytics. Tien (2013) emphasized the role these tools can have to aid in corporate decision making and how it can assist with understanding the vast quantities of data many corporations now gather.

Data which cannot be analyzed is of little benefit to, for example, a psychologist working in a corporate setting seeking to gather insights for an employer. Big data analytics gives businesses, as well as psychologists, the tools needed to extract useful information from otherwise potentially unusable sets of data. This use can be extrapolated to any group or organization that deals with vast amounts of data, such as governments or ambitious research projects.

“The Durkheim Project” is an ongoing program that uses big data technologies to tackle the issue of suicide in the veteran population (Patterns and Predictions, 2014). By monitoring the data from social networking accounts of United States military veterans big data analytics is used to predict the individual's risk of suicide in real-time. The hope of this project is that it could be used to offer intervention to those at risk individuals who may not be identified otherwise (Patterns and Predictions, 2014). With this project in mind it is easy to see the incredible potential for this technology and the opportunity it presents for clinical and preventative programs. This program requires minimal effort for participants and could become an easy way to unobtrusively monitor suicide risk. If it is successful it could be a great resource for all at-risk groups, not just veterans, as well as set a precedent for what researchers can do with social networking information and the potential help this kind of real-time monitoring can bring. Individuals have already incorporated social media into various aspects of their lives and, so long as it is done ethically, utilizing this information to help the user is a simple way to improve the experience (Oboler, Welsh, & Cruz, 2012). If conclusions can be drawn which help these individuals who consent to have their data analyzed, such as through “The Durkheim Project” or future programs which tackle other issues, then this could become an invaluable early warning system to catch and prevent all manner of problems. For example, a program that parents or children could opt into which would monitor potential cyber-bullying through social networks could be an invaluable way to tackle a tricky problem. Data from social networks offers up a multitude of insights into difficult problems which big data technologies could one day assist in understanding.

On the forensic side, one particular project seeks to combine information to form predictions that law enforcement can use to better focus efforts on high crime locations. The project uses historical, behavioural, and physical characteristics of communities to produce a digital map that models potential criminal activity and predicts future problem areas or trends for law enforcement (Collins, 2013). The physical characteristics refer to the map aspect of the program while the historical and behavioural side refers to information law enforcement gathers on crime rates, locations of crimes, and other pertinent data (Collins, 2013). Again, this is an ongoing pilot project but it is hoped that this process of “risk terrain modelling” (Caplan & Kennedy, 2011) for crime will yield new insights through the use of big data analytics into information law enforcement officials already have at their disposal. Using this tool could be highly beneficial for law enforcement and areas of forensic psychology involved in the predictive aspect of crime. If this were found to be successful on the larger scale, possibly a future project could focus on individual recidivism rates and what factors lead to criminal activities. This could aggregate information about the history of criminals, their behaviours, or other pertinent factors to help policy makers, law enforcement, and prisons make more informed decisions about what factors predict the likelihood of re-offending and how best to deal with this population.

These social science projects, for the most part, incorporate information that is already available but aggregate it in such a way that new insights, information, and predictions can be ascertained from the analysis of already existing data sets. It is through this process of better utilizing and connecting sources of available information that big data analytics can be put to great use. However, many researchers also use big data technologies to better organize and examine new data that has been gathered, such as “The Durkheim Project” (Patterns and

Predictions, 2014). There is a multitude of potential projects that could be undertaken using big data and similar methods to the projects going on currently if they are found to be successful. Better understanding consumer behaviour and concerns to meet customer's needs, social networking as an early warning system for users, and aggregating information about crime to form a more complete picture of various forensic problems are all potential applied uses for this technology. There are also benefits for more basic psychological research which could be achieved by looking at pre-existing sources of big data that cannot be analyzed through traditional means, or by artificially creating such a data set by forming one larger set of big data from a group of smaller subsets. By combining data sets and running analyses on one set of big data researchers could perform more diverse types of analyses and look at the research issue more broadly, hopefully achieving a type of “big picture” view of the problem or a more comprehensive look at the nuances of the data. Or, as previously stated, it need not rely on previously gathered data since this technology allows for new projects which deal with very large data sets gathered by the researcher with the intention of using big data analytics.

Psychological Well-Being

The current study will examine big data in psychology by looking at the potential predictors of psychological well-being in large data sets. This aspect of the study is simply an exercise used to evaluate big data technologies, their research applications, and potential benefits or deterrents that might exist for this type of research. The issue of psychological well-being is a tricky one which encompasses various factors that may not be prevalent or apparent. Psychological well-being is a multifaceted concept that refers to things such as mental health, anxiety or stress, happiness, self-esteem, and other related traits (James, Bore, & Zito, 2012). Research by Ryff (1995) demonstrates just how complex a theory of psychological well-being might be by incorporating several prominent theories and hypothesizing how psychological concepts like autonomy, environmental mastery, and several others might relate to well-being.

Other research found that there was a positive correlation between age, extraversion, education, and conscientiousness leading to better psychological well-being and better overall well-being (Keyes, Shmotkin, & Ryff, 2002). Studies have found many other correlations and predictors of psychological well-being as well, including everything from an individual's personality (James, Bore, & Zito, 2012) to their time spent engaging in leisure activities (Trainor, Delfabbro, Anderson, & Winefield, 2010) to their respective income within nations (Morrison, Tay, Diener, 2011). It is easy to see that a concept like psychological well-being could take many forms and be influenced by a plethora of factors. It is because of this that it would be beneficial to take a broad look at how various variables in an individual's life may be predictive of psychological well-being. Additionally, it would be helpful to see which variables are found to be significant when looking at a broad range of predictors. Analyzing psychological well-being through the lens of big data analytics may provide insight into relationships and variables that would otherwise be missed through conventional analysis.

Current Study

Real world issues are seldom ever as simple as they appear and often the underlying factors behind a result may themselves have underlying factors ad infinitum. It is for this reason that Schultes (2013) argued that variables which are assumed to be independent may not be and that subtle correlations exist throughout large data sets. He said that as the size of data sets increases it becomes easier to detect these small correlations, such that many small interconnections may exist in the case of vast amounts of data (Schultes, 2013). This tendency to stray away from rigid hypothesis driven big data research is also going on in biomedical literature. Miller (2012) explained that some researchers using big data analytics in medical databases have found a hypothesis-free approach to better suit the type of studies being conducted. By simply letting the technology run and seeing what results are obtained, researchers are finding information and connections which may have been missed and, once this is done, they can delve deeper into these findings without having to waste time on dead ends (Miller, 2012). Big data expands the scope of what researchers can look at, the methodology used to examine the issues, as well as how it can be looked at. Perhaps there are benefits to the type of exploratory research Miller (2012) described which can allow for previously unimagined or unseen interactions amongst the variables within large data sets. The current study will similarly attempt to take a broad exploratory approach to predictors of psychological well-being using big data technologies.

The current study will look at aggregating previous data to take an exploratory look at the issue of psychological well-being in Canadians through the use of big data analytics. Doing so will also serve as a testament to what kinds of basic research can be done with these tools and how psychology could utilize this technology to gather insights from vast data sets. Due to obvious limitations, this study will rely on previously collected data obtained from databases so measures of psychological well-being and potential predictors of it are dependent upon what measures were used in the previous research. For example, a hypothetical measure of psychological well-being could take the form of questions about stress levels or mental health concerns. Potential predictors of psychological well-being could take many forms like personality traits (James, Bore, & Zito, 2012), substance abuse, levels of leisure time (Trainor, Delfabbro, Anderson, & Winefield, 2010), or other factors not previously accounted for in smaller scale research.

The goal of the current study is more focused on the methodology of how research is done and the potential applications for big data analytics in psychology. This will involve learning about the various technologies and software used to analyze big data and discussing their potential benefits or limitations. As Ward (2013) stated there are some concerns with traditional statistical programs when looking at very large data sets, so it is useful to see how well this is mitigated by techniques and programs used to analyze big data. Because of this goal, predictors of psychological well-being are looked at to see what information can be extracted from the large data set using big data analytics. However, as was previously stated, this is more of an exercise used to demonstrate what big data analytics can do and an opportunity to evaluate these programs for psychological research. Additionally, depending on issues related to access and the current state of big data technologies, it is unknown how realistic working with big data analytics is for a typical untrained psychologist. However, assessing how realistic it is for social scientists to enter this field is primarily why the current study is being done. The main focus of this research is on the role big data may play in psychological research and how useful the current software and technology may be for this purpose. Looking into predictors of psychological well-being is, in and of itself, a meaningful task but it also serves to facilitate the main focus of the current study.

Data sets were collected, combined, organized, and analyzed to see what predictors of psychological well-being could be found. Attempts were also made to incorporate Hadoop based big data analytics programs into the project and evaluate them in regards to their perceived usefulness as a research tool. It is hypothesized that, by doing this, new insights could be gathered from the previously collected data into the issue of psychological well-being and the factors that influence it. Awareness will also be brought to what this technology is and its potential applications in the field of psychology. The awareness aspect will take the form of evaluating the current technologies available and learning how they can be applied to analyzing large data sets. The applications, limitations, or benefits of the current software used for big data analytics is outlined in relation to its potential uses in the field of psychology and, more broadly, the social sciences as a whole. Because of the dual nature of the current research project there exists two methods sections. The first methods section deals with the overarching exploration into big data technologies and the second section deals with the study examining predictors of psychological well-being.

Method

Participants

No participants were required for the exploration of the potential for big data technologies in psychology.

Materials

This project utilized a Windows 7 computer with a 3.50 GHz processor for running and evaluating the programs as well as extra RAM (16 GB total) for speed and additional hard drive space (763GB total) for running virtual systems. Other required materials included accessible software designed for big data, a stop watch for timing, and Statistics Canada data files (obtained from the Equinox data delivery system [part of the University of Western Ontario library system]).

Big data technologies used for comparative purposes included Hortonworks Sandbox 1.3 (running on a Linux based CentOS virtual machine), Cloudera Distribution Including Apache Hadoop 4.5.0 (running on a Linux based CentOS virtual machine), IBM InfoSphere BigInsights QuickStart 2.1 (running on a Linux based Red Hat virtual machine), and Microsoft's Windows Azure (an online cloud service including HDinsight, a Hadoop Distributed File System). All of these programs were freely available over the course of the project with the exception of Microsoft's Windows Azure where an academic pass was gifted for research purposes. Additionally, the IBM Statistical Package for the Social Sciences 20 (IBM SPSS Statistics 20) was used to test the baseline for what standard statistical programs are capable of.

Procedure

The first step was to design a sample study that could facilitate comparison between the big data and standard tools. This study took the form of looking for predictors of psychological well-being by amalgamating the Statistics Canada Canadian Community Health Survey (CCHS) data from 2003 to 2012 within SPSS along shared variables to create one large data set. Six different sets of data (CCHS 2004, CCHS 2005, CCHS 2007-2008, CCHS 2009-2010, CCHS 2011-2012) were included in this amalgamated set using SPSS' merge function. There were 141 variables shared amongst the iterations of the survey that made it into the amalgamated data set. The code names for these variables are included in Table 1 with the legend contained in the relevant Statistics Canada documents (Statistics Canada, 2013). Additionally, many variables had to be recoded for proper analysis to be done since the majority included non-answers, refusals, and non-applicable points numbered along the spectrum of participant responses. These were recoded to be system missing values within SPSS. Additionally, yes and no responses were reversed for directional consistency of the data. The full legend of recoded variables can be found in Appendix A.

Table 1: *List of CCHS shared variables by name.*

ADM_N09, ADM_N10, ADM_N11, ADM_PRX, ADM_RNO, ALC_1, ALC_2, ALC_3, CCC_071, CCC_101, CCC_121, CCC_131, CCC_141, CCC_171, CCCG102, DHH_OWN, DHH_SEX, DHHG611, DHHGAGE, DHHGHSZ, DHHGL12, DHHGLE5, DHHGLVG, DHHGMS, EDUDH04, EDUDR04, FVCDCAR, FVCDFRU, FVCDJUI, FVCDPOT, FVCDSAL, FVCDTOT, FVCDVEG, FVCGTOT, GEN_01, GEN_07, GEN_10, GEN_02B, GENDHDI, GENDMHI, GEODPMF, HWTGBMI, HWTGISW, HWTGWTK, INCG2, INCGHH, INCGPER, PAC_1A … PAC_1Z*, PAC_2A … PAC_2Z*, PAC_3A … PAC_3Z*, PACDEE, PACDFM, PACDFR, PACDPAI, PACFD, PACFLEI, SACDTOT, SDC_8, SDCFIMM, SDCGRES, SMK_10, SMK_202, SMK_204, SMK_01A, SMK_05B, SMK_05C, SMK_05D, SMK_06A, SMK_09A, SMK_10A, SMKDSTY, WTS_M

*Note: PAC_1, PAC_2, and PAC_3 each include all PAC variables from A to Z.

Ex., PAC_1A, PAC_1B, PAC_1C… PAC_1X, PAC_1Y, PAC_1Z.

The methodology for this sample study is listed within the Method (Study) section. Attempts were made to perform statistical analyses on this data set in big data technologies as well as the baseline SPSS analysis. Additionally, SPSS was evaluated for its ability to handle large data sets by timing basic multiple linear regression analysis on data exponentially increasing in size. This was done by taking the initial date file, merging cases with itself, and then saving the increased file and repeating the process with the increased file (effectively doubling the number of cases and size each time).

The next step was to find accessible big data analytics software, programs, or services based on Hadoop, the industry standard for big data, which would allow for analysis of a large data file as well as evaluation and comparison of the results of said analyses with the results of similar analyses conducted by a standard statistical tool used by psychologists. The benefit of Hadoop is that it reduces large jobs into smaller tasks, allocates those tasks, and allows multiple cores to work parallel to one another. This allows for the completion of larger tasks in a more realistic and timely manner than traditional programs. The step regarding finding these programs involved searching online for open-source software or low-cost programs and contacting corporations for evaluative versions of their software. Over the course of the project several Hadoop based programs were tested by attempting to perform similar statistical analyses within these big data analytics programs. This included testing Hortonworks Sandbox 1.3, Cloudera Distribution Including Apache Hadoop 4.5.0, IBM BigInsights QuickStart 2.1, and Microsoft's Windows Azure.

Method (Study)

Participants (Study)

This study included 681,578 participants from Canadian Community Health Surveys (archival Statistics Canada data). This includes 310,711 (45% of sample) males and 370,867 (55% of sample) females. Ages of participants ranged from 12 to over 80 (age was only recorded in brackets). The average age for male participants was within the 40–44 range while the average for female participants was within the 45-49 range. Individuals were contacted by Statistics Canada agents face-to-face through visits to homes, over the phone, and mailed letters for participation. The full method of recruitment for the Canadian Community Health Survey can be found in the relevant Statistics Canada documents (Statistics Canada, 2013).

Materials (Study)

The current study utilizes no materials aside from the statistical tools and technologies used for comparative purposes. However, materials required for the Statistics Canada data set include laptops, telephones, and the assisted interviewing software (Statistics Canada, 2013).

Procedure (Study)

Participants were visited at home by an interviewer who, upon finding a semi-private location, began asking them interview questions and recording answers on a laptop. Upon finishing participants would be thanked for their participation, have a chance to ask questions, and then their data would be sent in. Similarly, many participants had phone interviews which procedurally followed a similar format. Participants were called and interviewed while the interviewer recorded their responses on a computer. The full procedure for participants can be found in the relevant Statistics Canada documents (Statistics Canada, 2013).

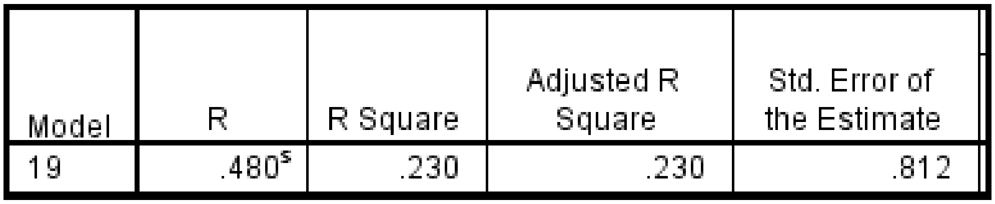
Results (Study)

Multiple linear regressions were conducted in SPSS 20 to determine possible predictors of psychological well-being which was operationally defined as the variable GENDMHI (self-reported mental health ranging from poor to excellent). Stepwise regression analysis was run with PIN of 0.05, a POUT of 0.10, and pairwise removal of missing values. These linear regressions were run on all variables as well as on subsets of the variables. From these analyses potential predictors to be included in a model of psychological well-being were found and a multiple regression was run on just the variables found to be significant predictors within the previous multiple linear regressions. The amount of variance accounted for by the best model can be found in Table 2. The regression equation for the best model is as follows:

GENDMHI = 1.231 + (.282)(GENDHDI) + (-.183)(GEN_07) + (.104)(GEN_10) +

(.077)(INCGPER) + (-.024)(DHHGHSZ) + (.012)(FVCDTOT) + (.168)(DHHGMS) + (-.059)(DHHGAGE) + (.131)(CCC_071) + (.047)(CCCG102) + (.037)(EDUDR04) + (.028)(ALC_3) + (.144)(CCC_121) + (.008)(HWTGBMI) + (.003)(PACDEE) + (.058)(DHH_SEX) + (.156)(CCC_131).

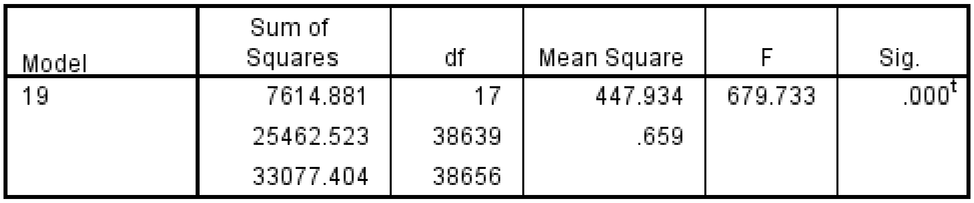
Table 2: *R and R Square Values for the Best Model*



Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
19	.480[s]	.230	.230	.812

A legend for each variable included in the best model can be found in Appendix B. This explains what each variable found to be a significant predictor of psychological well-being means. The correlations table and descriptive statistics for the best model can be found in Appendix C. Table 3 includes the ANOVA table of the best model.

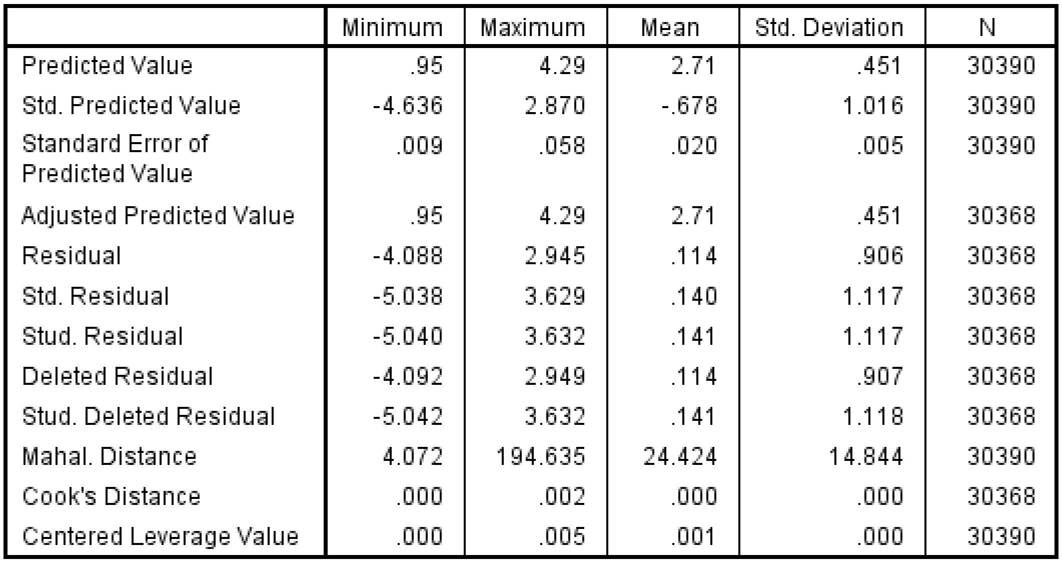
Table 3: *ANOVA Table for the Best Model*



Model	Sum of Squares	df	Mean Square	F	Sig.
19	7614.881	17	447.934	679.733	.000[t]
	25462.523	38639	.659		
	33077.404	38656			

The homogeneity of variances of the criterion variable for specific values of the predictors were also examined by using studentized residuals to determine whether or not the variances of the residuals differ. Also, Cook’s distances were used to check for the influence of outliers on the regression line. These results can be found in the residuals statistics within Table 4.

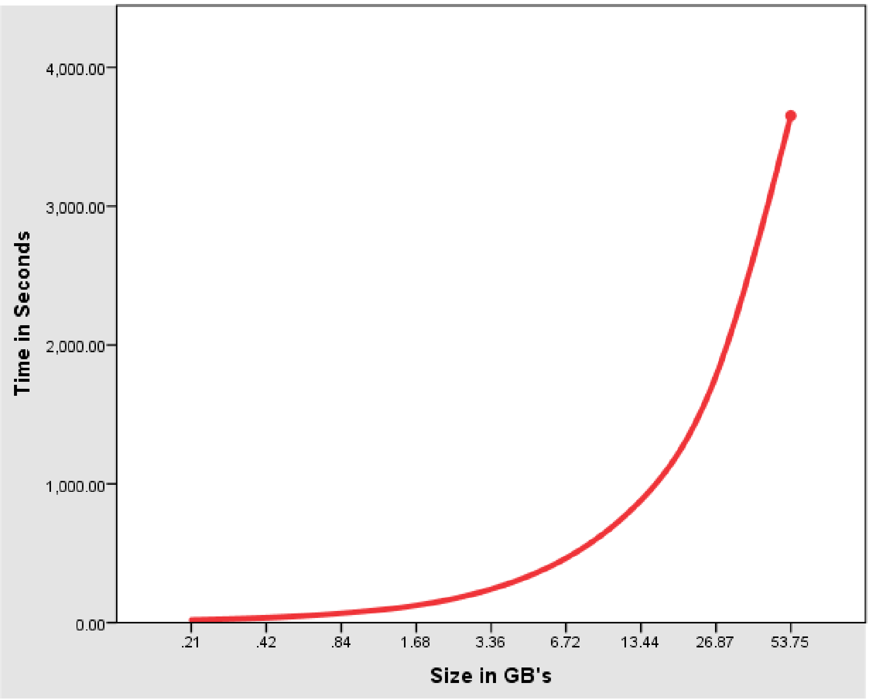
Table 4: *Residuals Statistics for Multiple Regressions Analysis*



	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	.95	4.29	2.71	.451	30390
Std. Predicted Value	-4.636	2.870	-.678	1.016	30390
Standard Error of Predicted Value	.009	.058	.020	.005	30390
Adjusted Predicted Value	.95	4.29	2.71	.451	30368
Residual	-4.088	2.945	.114	.906	30368
Std. Residual	-5.038	3.629	.140	1.117	30368
Stud. Residual	-5.040	3.632	.141	1.117	30368
Deleted Residual	-4.092	2.949	.114	.907	30368
Stud. Deleted Residual	-5.042	3.632	.141	1.118	30368
Mahal. Distance	4.072	194.635	24.424	14.844	30390
Cook's Distance	.000	.002	.000	.000	30368
Centered Leverage Value	.000	.005	.001	.000	30390

The multiple regression analyses that were timed in SPSS 20 show a distinct trend in the amount of time it takes to process larger data sets. The original data file used in the study on psychological well-being was .21GB in size and analysis took 20 seconds to return output. After doubling the number of cases in the data set eight times and timing how long it took SPSS 20 to produce output for each iteration it was found that a 53.75GB file took over an hour (3,852 seconds) to analyze. It is important to note that, for SPSS 20, analysis is performed on the King's University College computer servers as opposed to the local machine. The results from the sizing tests can be found in Table 5.

Table 5: *Time Spent Producing Output as a Result of File Size*



Results

The exploration into big data technologies ran into several issues over the course of the study. Unfortunately, none of the big data analytics programs or services tested were able to properly provide the level of analysis expected for this kind of research.

Hortonworks Sandbox 1.3 has a relatively simple set up, did not require resource intensive hardware, and the lessons and tutorials built-into the program were useful. However, the Sandbox version of Hortonworks is a learning tool and not capable of full-fledged big data analytics in addition to having no statistical analysis available. Several tutorials were completed and some basic commands run but, after doing so, it was decided that this program was unsuitable for the needs of the study.

The next program tested was the Cloudera Distribution Including Apache Hadoop 4.5.0 which, when used, was only a free version that lacked the functionality of the full version. Additionally, components of the software came deactivated, error messages were present, and other configuration issues marred the experience as well as limited what could be done. After attempting and failing to fix the issues that were present upon a fresh installation of the program it was decided that further options would be looked at.

IBM InfoSphere BigInsights QuickStart 2.1 was the best in terms of functionality and usability. It allowed the data to be easily displayed and appeared to have statistical analysis present in some form. However, this had the greatest hardware requirements of the programs used which led to issues and slow down during resource intensive commands or applications. Additionally, while the statistical program R (a popular open source statistics program) was present in some form (there was an application within IBM InfoSphere BigInsights QuickStart 2.1 that could access it) actually implementing it appeared to be impossible given that the program had stopped being updated for the Linux based Red Hat operating system that BigInsights was run on. Activating statistical analysis, while hypothetically possible, was not feasible. Also, on several occasions within multiple data sets the program was found to drop cases which caused concern. A data file was uploaded and some basic Pig (an application within Hadoop used to issue commands and interact with data) commands run by the researchers but because of the issues with speed and statistical analysis this program did not fulfill the expectations of the current project. Additionally, the researchers requested access to the IBM SPSS Analytic Catalyst, a version of the standard statistical software that could be integrated into BigInsights and big data research, which would have solved some of the issues with a lack of statistical analysis but were ultimately denied access.

The last big data technology utilized was Microsoft’s Windows Azure which was the only online cloud based Hadoop system used. While this did bypass many of the hardware and configuration issues present with the other programs it had problems of its own. Windows Azure is a conglomerate of various products and services of which big data analytics is one component. The website itself was not user friendly and was more akin to a status window which gave information on what services were implemented, what files were stored, and allowed the user to create or check Hadoop clusters. However, any actual functionality in regards to uploading data, interacting with data, viewing data, or analyzing data was detached from the website. Commands had to be run through a separate downloadable command line interface that operated on a local machine. This interface offered little help to a user looking to learn how to interact with the data and appeared quite crude when compared to the other big data technologies evaluated. Additionally, the only way statistical analysis could be run on the data set was by downloading the file and having it locally analyzed by Excel, thereby defeating the purpose of a cloud computation service. A data file was uploaded into Windows Azure and attempts were made to connect Excel to perform statistical analysis but this proved unsuccessful due to errors in the program. Due to the crude interface, lack of cloud computation for statistical analysis, and issues present this service failed to meet the standards required for the study. Time constraints prevented additional big data technology options from being explored.

Discussion

The study into predictors of psychological well-being allowed for an interesting comparison of many diverse variables that are not normally analyzed together. While the strongest predictors of psychological well-being found have appeared in the previous literature, such as youth, education (Keyes, Shmotkin, & Ryff, 2002), and income (Morrison, Tay, Diener, 2011), using a large data set and not coming in with preconceived notions allowed insights that may have been missed or otherwise deemed insignificant. For example, work by Gestel, Jansen, and Theunissen (2011) found that binge drinking was a predictor of poorer mental health outcomes in young Dutch teens (12 – 15) but not older teens (16 – 18). In the current study ALC_03 (binge drinking) was found to have a very low effect size (r = -.007, p < .000) which may demonstrate that the research by Gestel, Jansen, and Theunissen (2011) is more indicative of a small but real trend and may hold true for a broader group than originally studied. However, more complex analysis performed on the Canadian Community Health Survey data set could reveal whether or not this correlation only holds true for certain ages. Without a fairly large data set it is likely that this effect may have been missed. Similarly, research looking at diet and mental health by Cook and Benton (1993) found a positive correlation between fruit and vegetable intake akin to what was found in the current study, but only for females. The current study found a general effect (*r* = .098, *p* < .000) so further research looking into these correlations within this data set could tease apart whether this relationship changes when gender is factored in (or age in the case of the binge drinking findings). Looking at these relationships through the context of a large data set allows for more diverse findings and a way to discover areas that could merit further research into why particular relationships exist and in what circumstances.

Some predictors of psychological well-being were found that may have been missed in the current literature. Some of these unique findings may include household size (*r* = .081, *p* < .000), heart disease (*r* = -.071, *p* < .000), and high blood pressure (*r* = -.074, *p* < .000). This may be due to these variables not being examined, having small effects that did not show up in a smaller data set, or a multitude of other reasons. This is one of the benefits of working with very large samples since it allows the data to speak for itself. However, it also calls into question the manner by which psychologists and other scientists determine what exactly makes a relationship significant. With a large enough sample even incredibly small effects become significant and further consideration must be given as to whether or not a significant effect is truly a meaningful relationship. With this new trend towards big data research a conversation will have to be had regarding what constitutes a real effect due to the significance of, conceivably, nearly every variable with a large enough sample size. However, even small effects can be important when it comes to, for example, drug interactions which may be lethal for small populations (Miller, 2012). This topic may become a growing problem in coming years that warrants further discussion and research.

Big data research, while potentially difficult for psychologists given the current climate, is still a viable research opportunity with a myriad of benefits. There is potential for this technology to benefit the social sciences by expanding what populations can be examined, what solutions can be created, and what can be learnt from already existing data. The current deterrents for social scientists are things that can be and are being solved by improvements to technology. It is likely that many of the problems that exist will be solved by technological advancements and the proliferation of this software into broader areas. The main issues that are holding back this type research, as identified by the current study, fall within four general areas of concern.

The first is issues related to accessibility and cost. This project experienced a multitude of stumbling blocks when trying to procure big data technologies for free or minimal costs that contained the functionality required for psychological research. One of the only ways to access the software, initially, was to contact representatives from the companies creating these technologies and ask them for software. This was problematic since many were hesitant to provide data for independent researchers performing a, relatively, small project. However, over the course of the study this did improve and more avenues, such as academic passes and application programs, were developed that allowed for easier access. Additionally, many extra resources not commonly deployed in these programs are needed for research of this nature, such as statistical programs that work in-conjunction with big data technologies which require additional costs that would be unreasonable for a lone psychologist to incur.

The next issue deals with the lack of statistical analysis integrated into the majority of these programs. Many simply, during the time of the study, did not have available the types of analysis required for psychological research built-in or easily accessible. All of the programs that were tested in the current study lacked support for performing the kinds of statistical analysis commonplace in psychological research. This is a major deterrent since it requires data to be exported into another program which, depending on the size or nature of the data set, would be subject to the limitations present in the standard statistical programs.

Another potential issue may be the requisite technological knowledge required for properly using big data technologies. Many required a basic to moderate understanding of programming, running virtual machines, and knowing how to navigate non-typical operating systems. All of the current programs tested required the writing of commands in various languages (Pig Latin, HiveQL, Java, R, etc.) and the majority had to be run as a Linux virtual machine which may be difficult for users to set up or navigate without prior experience.

Finally, the last issue was in regards to the hardware needed to run big data technologies. Many of the programs require computers more powerful than what a typical social scientist may have access to, a network of computers, or cloud computation. For example, IBM InfoSphere BigInsights QuickStart 2.1 required a minimum of 8GB of RAM allocated to the virtual machine in order to run. Even with this allocation it was slow and problematic until more RAM was added. This, in addition to the other issues, may deter psychologists and partially explain why this area of research is currently not being utilized.

The timed regression speed tests performed in SPSS demonstrate that current statistical software has the capability to analyze traditional data sets and even quite large (by psychology standards) data sets. This software will likely suffice for the bulk of research in psychology and the social sciences. However, there is more that must be accounted for. Increasingly complex forms of analysis and actual big data greatly increase the time it takes to run analyses potentially making it unreasonable or even undoable. The largest file attempted was 54GB for which the analysis took over an hour. Big data is typically considered to be within the 1,000GB to

1,000,000GB range (Tien, 2013) so one can imagine how much longer even the most basic forms of analysis could potentially take. Additionally, non-traditional data sets which include a variety of data types or are unstructured as well as real-time data flows still present problems for traditional statistical analysis programs that are not present in big data analytics. This is one potential area that can be aided by more integration of this technology in the social sciences. By getting the perspective of psychologists and social scientists some of the issues that may not be noticed in industry could be identified and solutions developed.

Due to the time and accessibility limitations within the current study there is still a lot of future research that should be done on the viability of big data technologies as well as research with big data within psychology and the social sciences. Potential projects could evaluate all the leading statistical analysis programs normally utilized within the social sciences to see if any outperform others in terms of large data or big data capabilities. SPSS, SAS, Stata, R, Excel, and other leading analysis software could be tested and evaluated on their ability to do this kind of research similar to what was done by Ward (2013) but with a greater focus on big data analytics.

Another area that could be examined is comparisons of standard statistical software and their emerging big data counterparts. SPSS and SAS have both recently released versions of their normal statistical analysis software that has been designed for use in conjunction with big data technologies. Comparing the functions present in these new versions with what can be done in the traditional versions would be useful as well as evaluating how user friendly these big data programs are and whether or not additional knowledge is needed above and beyond knowledge of how to use the traditional statistical programs. These new technologies have the potential to solve a lot of the issues currently present in big data analytics software (lack of statistical integration, not accessible without requisite knowledge, etc.) and may be what is needed to attract more psychologists or other social scientists.

The current study, as previously mentioned, only looked at one of the V's of big data (Deroos et al., 2012) so further projects could examine how standard statistical software deals with the other two. Since a large volume of data was easiest to access this is what was looked at within the current study. Future research could evaluate how variety and velocity are dealt with in big data analytics as well as traditional statistical software. This may be a shortcoming of the current traditional statistics software and warrants further investigation. Comparing functionality to what is offered by big data analytics programs or even the newly released big data versions of traditional statistical software would be a useful endeavour. Perhaps big data technologies and big data statistical software would be more capable of examining data sets where variety and velocity are present.

References

Caplan, J. M., & Kennedy, L. W. (Eds.). (2011). *Risk terrain modeling compendium.* Newark, NJ: Rutgers Center on Public Security.

Collins, H. (2013). Predicting crime using analytics and big data. *McClatchy - Tribune Business News*

Cook, R., & Benton, D. (1993). The relationship between diet and mental health. *Personality and Individual Differences,14*(3), 397-403. doi:10.1016/0191-8869(93)90308-P

Davenport, T. H., & Barth, P. (2012). How big data is different. *MIT Sloan Management Review, 54*(1), 43-46.

Deroos, D., Deutsch, T., Eaton, C., Lapis, G., & Zikopoulos, P. (2012). *Understanding big data: analytics for enterprise class Hadoop and streaming data*. New York: McGraw-Hill.

Gestel, A., Jansen, M., & Theunissen, M., (2011). Are mental health and binge drinking associated in Dutch adolescents? Cross-sectional public health study. *BMC Research Notes*, *4*(1), 100-100. doi:10.1186/1756-0500-4-100

James, C., Bore, M., & Zito, S. (2012). Emotional intelligence and personality as predictors of psychological well-being. *Journal of Psychoeducational Assessment, 30*(4), 425-438. doi:10.1177/0734282912449448

Keyes, C. L. M., Shmotkin, D., & Ryff, C. D. (2002). Optimizing well-being: The empirical encounter of two traditions. *Journal of Personality and Social Psychology, 82*(6), 1007-1022. doi:10.1037/0022-3514.82.6.1007

Koh, J. (2013, Jun 18). Understanding consumers with help from big data. *The Business Times*. Retrieved from https://www.lib.uwo.ca/cgibin/ezpauthn.cgi/docview/1368686527?accountid=15115

Lesk, M. (2013). Big data, big brother, big money. *IEEE Security & Privacy, 11*(4), 85-89. doi:10.1109/MSP.2013.81

Miller, K. (2012, January 2). Big data analytics in biomedical research. *Biomedical Computation Review*. Retrieved November 10, 2013, from http://biomedicalcomputationreview.org/content/big-data-analytics-biomedical-research

Morrison, M., Tay, L., & Diener, E. (2011). Subjective well-being and national satisfaction: Findings from a worldwide survey. *Psychological Science*, *22*(2), 166-171. doi:10.1177/0956797610396224

Oboler, A., Welsh, K., & Cruz, L. (2012). The danger of big data: Social media as computational social science. *First Monday, 17*(7) doi:10.5210/fm.v17i7.3993

Ovadia, S. (2013). The role of big data in the social sciences. *Behavioral & Social Sciences Librarian*, *32*(2), 130-134.

Pandiani, J. A., & Banks, S. M. (2003). Large data sets are powerful. *Psychiatric Services (Washington, D.C.)*, *54*(5), 745; author reply 746-746. doi:10.1176/appi.ps.54.5.746

Patterns and Predictions (2014). Automated flagging for psychological health. *Durkheim Project*. Retrieved March 26, 2014, from http://durkheimproject.org/

Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology, 57*(6), 1069-1081. doi:10.1037/0022-3514.57.6.1069

Ryff, C. D. (1995). Psychological well-being in adult life. *Current Directions in Psychological Science*, *4*(4), 99-104. doi:10.1111/1467-8721.ep10772395

Schadt, E. E. (2012). The changing privacy landscape in the era of big data. *Molecular Systems Biology*, *8*, 612. doi:10.1038/msb.2012.47

Schultes, E. (2013, June). Big data: The myth of independent variables. *The 32nd Annual Conference of the Society for Scientific Exploration*. Lecture conducted from Dearborn, MI.

Statistics Canada (2013). *Canadian Community Health Survey (CCHS) annual component: User guide 2012 and 2011-2012 microdata files*. Retrieved from: http://equinox2.uwo.ca/docfiles/cchs/2012/cchs-escc2012_2011-2012gid-eng.pdf.

Tien, J. M. (2013). Big data: Unleashing information. *Journal of Systems Science and Systems Engineering*, *22*(2), 127-151. doi:10.1007/s11518-013-5219-4

Trainor, S., Delfabbro, P., Anderson, S., & Winefield, A. (2012). Leisure activities and adolescent psychological well-being. *Journal of Adolescence*, *35*(2), 467-467. doi:10.1016/j.adolescence.2012.02.005

Ward, B. W. (2013). What’s Better—R, SAS®, SPSS®, or stata®? Thoughts for instructors of statistics and research methods courses. *Journal of Applied Social Science*, *7*(1), 115-120. oi:10.1177/193672441

Appendix A

RECODING

1 = 2, 2 = 1, 6 7 8 9 = 999

ADM_PRX
ALC_1
CCC_071
CCC_101
CCC_121
CCC_131
CCC_141
CCC_171
PAC_1A – PAC_1Z
PACFLEI
SDCFIMM
SMK_10
SMK_05D
SMK_06A
SMK_09A
SMK_10A

96 97 98 99 = 999

CCCG102
DHHGLVG
SACDTOT
SMKDSTY

99.9 = 999

PACDEE

96 = 0, 97 98 99 = 999

ALC_2
SMK_05C

96 = 1, 97 98 99 = 999

ALC_3
INCGPER

996 = 0

PAC_2A – Z
SMK_204
SMK_05B

6 7 8 9 = 999

ADM_N09
ADM_N10
DHH_OWN
DHHGHSZ
DHHGMS **(Also, recoded values into not married [1] and married [2])**
EDUDH04
EDUDR04
FVCGTOT
GEN_01
GEN_02B
GEN_07
GEN_10 **(Also, values reversed for consistency [ex., 1 = Very Weak])**
GENDHDI
GENDMHI
HWTGISW
INCG2
INCGHH
PACDFR
PACDPAI
PACFD
SDCGRES
SMK_202
SMK_01A

6 = 1, 7 8 9 = 999

ADM_N11

6 = 0, 7 8 9 = 999

PAC_3A – Z

1 = 2, 2 = 1, 6 = 1, 7 8 9 = 999

SDC_8

990 – 1000 = SYSMIS

All variables

Appendix B

ALC_3 (*Binge Drinking*): Frequency of having 5 or more drinks in past year.

CCC_071 (*High Blood Pressure*): Yes or no response for presence of medical condition.

CCC_121 (*Heart Disease*): Yes or no response for presence of medical condition.

CCC_131 (*Cancer*): Yes or no response for presence of medical condition.

CCCG102 (*Onset of Diabetes*): Indicates age participant was when diagnosed with diabetes.

DHHGAGE (*Age*): Indicates the age bracket participant falls within.

DHHGHSZ (*Household Size*): Indicates how many people live with the participant.

DHHGMS (*Marital Status*): Indicates if participant is married or single.

DHHSEX (*Gender*): Indicates if participant is male or female.

EDUDR04 (*Education Completed*): Indicates highest level of education completed.

FVCDTOT (*Fruit & Vegetable Intake*): Frequency of eating fruits and vegetables daily.

GEN_07 (*Life Stress*): Indicates perceptions of stress present in participant’s life.

GEN_10 (*Community Belonging*): Indicates participant’s sense of belonging to a community.

GENDHDI (*Self-Report Physical Health*): General physical health from poor to excellent.

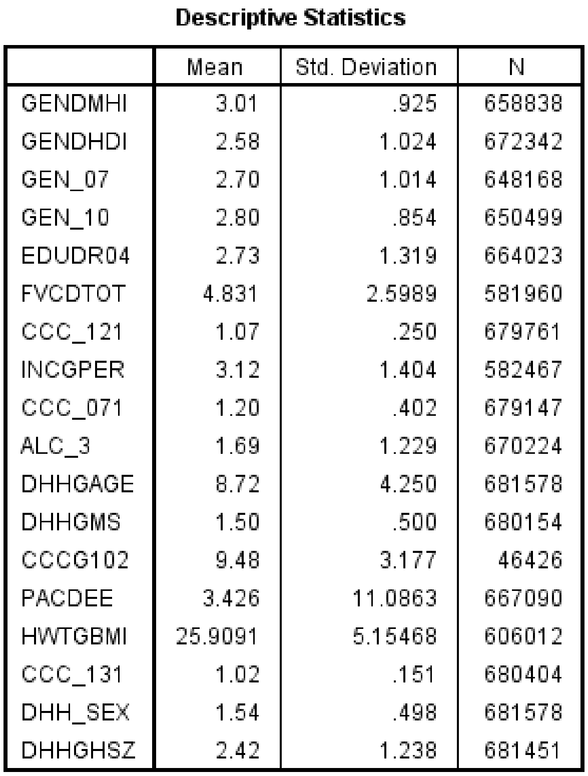
GENDMHI (*Self-Report Mental Health*): General mental health from poor to excellent.

HWTGBMI (*Body Mass Index*): Self-reported BMI score.

INCGPER (*Personal Income*): Reported personal income from all sources.

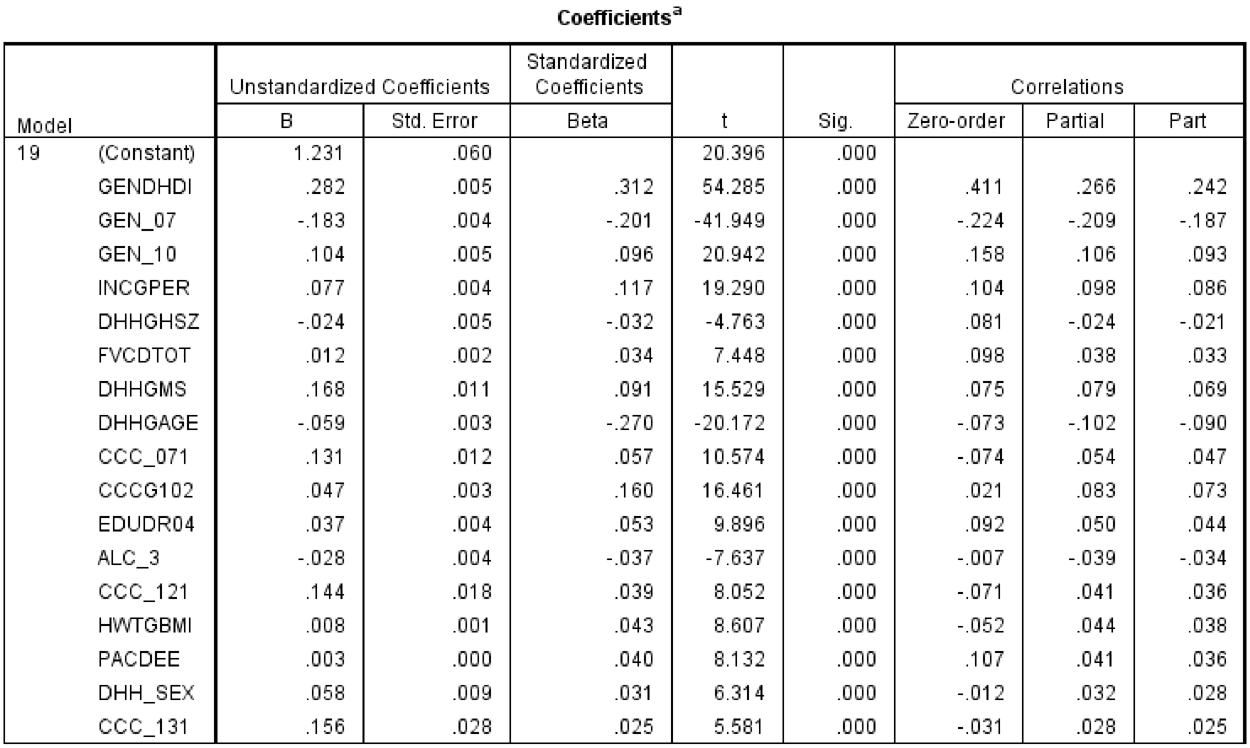
PACDEE (*Energy Expenditure*): Energy spent on physical leisure activities daily.

Appendix C



Descriptive Statistics

	Mean	Std. Deviation	N
GENDMHI	3.01	.925	658838
GENDHDI	2.58	1.024	672342
GEN_07	2.70	1.014	648168
GEN_10	2.80	.854	650499
EDUDR04	2.73	1.319	664023
FVCDTOT	4.831	2.5989	581960
CCC_121	1.07	.250	679761
INCGPER	3.12	1.404	582467
CCC_071	1.20	.402	679147
ALC_3	1.69	1.229	670224
DHHGAGE	8.72	4.250	681578
DHHGMS	1.50	.500	680154
CCCG102	9.48	3.177	46426
PACDEE	3.426	11.0863	667090
HWTGBMI	25.9091	5.15468	606012
CCC_131	1.02	.151	680404
DHH_SEX	1.54	.498	681578
DHHGHSZ	2.42	1.238	681451



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
19	(Constant)	1.231	.060		20.396	.000			
	GENDHDI	.282	.005	.312	54.285	.000	.411	.266	.242
	GEN_07	-.183	.004	-.201	-41.949	.000	-.224	-.209	-.187
	GEN_10	.104	.005	.096	20.942	.000	.158	.106	.093
	INCGPER	.077	.004	.117	19.290	.000	.104	.098	.086
	DHHGHSZ	-.024	.005	-.032	-4.763	.000	.081	-.024	-.021
	FVCDTOT	.012	.002	.034	7.448	.000	.098	.038	.033
	DHHGMS	.168	.011	.091	15.529	.000	.075	.079	.069
	DHHGAGE	-.059	.003	-.270	-20.172	.000	-.073	-.102	-.090
	CCC_071	.131	.012	.057	10.574	.000	-.074	.054	.047
	CCCG102	.047	.003	.160	16.461	.000	.021	.083	.073
	EDUDR04	.037	.004	.053	9.896	.000	.092	.050	.044
	ALC_3	-.028	.004	-.037	-7.637	.000	-.007	-.039	-.034
	CCC_121	.144	.018	.039	8.052	.000	-.071	.041	.036
	HWTGBMI	.008	.001	.043	8.607	.000	-.052	.044	.038
	PACDEE	.003	.000	.040	8.132	.000	.107	.041	.036
	DHH_SEX	.058	.009	.031	6.314	.000	-.012	.032	.028
	CCC_131	.156	.028	.025	5.581	.000	-.031	.028	.025