International Journal of Business and Applied Social Science (IJBASS)

VOL: 4, ISSUE: 12
December/2018
https://ijbassnet.com/
E-ISSN: 2469-6501

BIG DATA ANALYTICS OF LABOR COST OVER THIRTY YEARS

MOUSUMI BHATTACHARYA
Charles F. Dolan School of Business
Fairfield University
Fairfield, CT 06824
Tel: (203) 254-4000 ext.2893
Fax: (203) 254-4105
E-mail: mbhattac@fairfield.edu
USA

ABSTRACT

Businesses need to analyze labor cost so that they understand what components of businesses affect the amount of resources spent on employees. Interpreting this will help make managerial decisions in adjusting costs of production and develop new strategies regarding investing in workers. Labor cost can vary due to many factors such as employee numbers, assets, liabilities, sales, and debt. In this study, I analyzed labor cost in relation to other firm characteristics, using a large panel data set over thirty years with over 24,000 firm-year observations. I used two different statistical software, R and SPSS and the results are compared. Results show that they both give the same output when running a multivariate regression. Both software are powerful enough to analyze large datasets; however the form of input data and treatment of missing values, matter in determining which one is more efficient. Managers and practitioners can use business analytics with big data to draw conclusions and make important managerial decisions. Future study should include big data analysis using more complex analytical techniques.

Keywords: HR Analytics, Labor cost

Acknowledgement: I want to thank Siddharth Jain, Research Assistant for his help in data analysis.

INTRODUCTION

The application of big data analysis in human resource (HR) management is currently at its early stage. Big data analytics is growing fast as organizations are beginning to leverage these to gain competitive advantage (Grover and Kar,2017). This paper examines the relevance of analyses of big data on labor cost and whether different types of software make a difference in the analytics. I applied R software and SPSS statistical package to a panel and cross-sectional data set on companies from the COMPUSTAT North America database from Wharton Research Data Services (WRDS) over the past thirty years. I ran multivariate regressions on labor cost and labor cost variability, as related to several other variables, to compare the results of the two statistical software. Although the statistical results where similar, we found differences in the process of computing, which might affect the choice of analytics software and technique.

Big data analytics offers various benefits to the organizations, by providing visual tools and multiple-loop analysis to give fine-grained results that enhance the quality of decisions taken (Li, Tao, Cheng, and Zhao, 2015). So far, the issue is that statistical packages like SPSS may not be able to find fine-grained relationships in the data because it analyses data using single-loop modeling. This means that it develops relationships between variables in the form of mathematical equations. In contrast, R software uses machine learning techniques to perform the statistical analyses. This is a significant research area because machine learning is an algorithm that can learn from data without relying on rules-based programming. It can detect smaller relationships between variables, which can help make many decisions in management, health sciences, cyber security and other areas.

This research contributes to the field of HR analytics study because, through big data analysis, it provides light on the relationship between organizational level factors and labor cost. The study highlights the importance of allocating resources to certain aspects of a company for competitive advantage. At the same time this study demonstrates the relative merits of R software and SPSS statistical package. In the future, researchers and practitioners can use this methodology and analytics to decide on business strategy for the success of their company.

THEORY DEVELOPMENT

Labor Cost

Labor is a critical component of businesses and a significant part of production input costs (Blinder, 1990; Freeland, Anderson, and Schendler 1979). Callahan et al. (2010) found that labor leverage, which is the ratio of fixed costs to variable costs of a company, is positively associated with the implied cost of capital. Another study conducted in Serbia by Kljenak, Radojko, and Jovancedvic (2015) analyzed labor costs in the trade market. They concluded that there is a significant participation of labor costs in the total cost of firms around the world. For example,in the retail industry, labor cost is 11.4% of total cost in Canada, 22.5% in Australia, 15.5% in UK, 24% in USA, and 27% in Russia (numbers from 2011-14). Their study showed that there is a significant positive relationship between labor costs and operating income in Serbia. Similar results were found when the study was extended to the UK and Australia. The study concluded that the number of employees and labor costs both indicate a company's performance.

Human capital assets are positively associated with analytics forecast long-term growth rates (Ballester et al., 2002). There was a significant positive relationship between human capital and forecasts for long-term growth. Labor costs were also positively related with industry adjusted average salary and industry concentration ratio. A similar study analyzing the effects of human resource management on small firms' productivity and employees' wages concluded that pharmacies could gain from aligning their wage policies with employees' contribution to firm performance (de Grip, and Sieben, 2005).

Given the significance of labor cost in a firm's resource allocation and performance, it is critical that HR analytics focus the management of labor cost in relation to other resources. Labor costs are composed of many subcomponents such as salaries, benefits, pensions, and profit sharing. The amount of resources allotted for labor costs varies based on the firm's total profit as well as the amount of resources allocated for other aspects, such as capital expenditure, assets, debt, liabilities and sales. In this paper we analyze several such relationships through multivariate regression analysis. This research can help managers balance where resources are allocated for overall success of the business.

Business Analytics and Big Data

More organizations are storing, processing, and extracting value from data of all forms and sizes. Big Data is a term that describes the large volume of data both structured and unstructured that inundates a business on a day-to-day basis. Big data is defined as large data sets that have more varied and complex structure (Jain et al., 2016). However the amount of data is not as important to an organization as the analytics that accompany it. When companies analyze Big Data, they are using business analytics to get the insights required for making better business decisions and strategic moves. Data analytics involves the process of researching big data in order to reveal hidden patterns, which are unable to be easily detected with other methods. Business Analytics is the study of data through statistical and operations analysis, the formation of predictive models, application of optimization techniques, and the communication of these results to customers, business partners. Business Analytics requires quantitative methods and evidence-based data for business modeling and decision making; as such, Business Analytics requires the use of Big Data.

Data analytics is the process of using structured and unstructured data through various analytical techniques. Machine learning is becoming popular with data analytics, especially big data because the machine does the difficult computations for us (Marsland,2011). For example, in predictive analysis, such as regression, the purpose is to predict the value of a particular variable (target or dependent variable) based on values of some other variables (independent or explanatory variables). In machine learning, regression is an example of **supervised** **l**earning because we are telling the algorithm what to predict. Machine learning models gain knowledge from existing patterns in data, teach itself and apply what has been learned to make future predictions.As more data become available, the machine learns from forecasting successes and failures and then updates predictive algorithms accordingly (Grable, and Lyons, 2018).

R and SPSS

R is a package-based language that uses the machine learning technique of computation. For almost all statistical techniques, the chances are that a package exists. It allows a great variety of analysis to be done from one source. R has a wide range of uses and can be applied to a large variety of tests (Park, 2009). Due to the machine learning in R, it should be able to read unstructured data and detect smaller relationships between the variables. Through training modules R alters testing methods in order to better fit the data to the model. R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including Linux), Windows and MacOS.

SPSS Statistics is a software package, owned by IBM corporation used statistical analysis. SPSS Statistics simplifies, or often do not need programming. SPSS datasets have a two-dimensional table structure, where the rows typically represent cases (such as individuals or organizations) and the columns represent measurements (such as age, income). Only two data types are defined: numeric and text (or "string"). All data processing occurs sequentially case-by-case through the file (dataset).

Missing values

A common task in data analysis is dealing with missing values. When the dataset is not complete, and when information is not available we call it *missing values*. Big data in labor cost typically has a large number of missing values because labor cost is not a required disclosure in 10k filings. Therefore researchers need to devote substantial effort in data cleaning (Osborne 2013). The need for and approach to data cleaning can vary widely depending on the software used for data analyses (Liu-Thompkins, and Malthouse,2017). Missing data can be a problematic when analyzing big data. If the amount of missing data is very small relatively to the size of the dataset, then leaving out the few samples with missing features may be the best strategy in order not to bias the analysis. However leaving out available data points deprives the data of some amount of information and depending on the situation, a data analysts may want to look for fixing the data before wiping out potentially useful data points from the dataset.

R and SPSS deal with missing data differently. In R, missing cells must be replaced with NA. Unlike SPSS, R uses the same symbol for character and numeric data. To identify missing values in your dataset the function is.na().When running functions, one parameter that must be dealt with is the NA fields. It can be omitted, excluded, passed, and more. Most modeling functions in R offer options for dealing with missing values. One can go beyond pair wise or list wise deletion of missing values through methods such as multiple imputation. The mice (multivariate imputation by chained equations) package in R, helps impute missing values with plausible data values (van Buuren, and Groothuis-Oudshoorn, 2011). These plausible values are drawn from a distribution specifically designed for each missing datapoint, thereby keeping the distribution of data the same (van Buuren, and Groothuis-Oudshoorn, 2011). Therefore R provides lots of options in handling missing values.

However, if NA values are passed into SPSS, it will read it as a string, rather than numeric values. Missing data must have blank cells and SPSS will replace it with a period, indicating that it is a missing value. SPSS analysis commands that perform computations handle missing data by omitting the missing values. For descriptive analysis for each variable, the number of non-missing values is used. You can specify the missing=list wise subcommand to exclude data if there is a missing value on any variable in the list. By default, correlations are computed based on the number of pairs with non-missing data (pair wise deletion of missing data). The missing=list wise subcommand can be used on the corr command to request that correlations be computed only on observations with complete valid data for all variables on the var subcommand (list wise deletion of missing data). In regression if values of any of the variables on the var subcommand are missing, the entire case is excluded from the analysis (i.e., list wise deletion of missing data). It is possible to further control the treatment of missing data with the missing subcommand and one of the following keywords: pair wise, mean substitution, or include. However, generally speaking, SPSS has less flexibility in handling missing values in data.

METHODOLOGY

I used panel data on firms for the past thirty years. Firm-year level data were extracted from COMPISTAT based on the condition that labor cost data is available. Each firm has at least five years of data, but most firms have several years of data. The panel data will provide information regarding relationship between various components of firms. Panel data is multi-dimensional data involving measurements over time. Panel data contains observations of multiple variables over a time period for the same firms. The data set contained organizational level factors for 1,212 companies over multiple years. See Table 1 for the list of variables included in the study. The database contains U.S. and Canadian fundamental and market information on more than 24,000 active and inactive publicly held companies. The data is analyzed using RStudio on OSX, as well as SPSS statistical package.

Sample

Data for the study was taken from COMPUSTAT North America database from the Wharton Research Data Services (WRDS) spanning about thirty years for some companies. Long time spans were needed to ensure that trends are significant and unbiased. There were 24,186 firm-year observations in the dataset. Many of the variables were skewed due to the large range of different firms (see histograms in Figures 1A-H). The natural logarithm was taken to normalize the data. Labor cost (XLR), number of employees (EMP), assets total (AT), liability total (LT), current assets (CA), current liabilities (CL), capital expenditure (CAPEX), net sales (SALE), cost of goods sold (COGS), selling and general administrative expenses (XSGA), operating income after depreciation (OIADP), cash flow (CFL), debt To assets, current ratio (CR), and quick ratio (QR) all had to be normalized. Working capital (WC), working capital turnover (WCTO), cash flow margin (CFLM), and degree of operating leverage (DOL) already were in a normal distribution. Dol Dummy was a bimodal distribution

Correlations

In R, in order to run the correlation matrix, the .csv file of the data first had to be imported into RStudio. After creating an object for the data, the database then had to be attached to the R search path allowing objects in the database to be easily accessed by their name. To run a correlation matrix using the build in r function cor(), I put parameters of the data object, and to use complete.obs. Complete.obs ignores any missing data in the dataset. However, to receive more information, I installed the Hmisc package Version 4.1-1 from cran.r-package.org. Using the rcorr() function in Hmisc and giving it a parameter of the dataset object as a matrix, RStudio gave the correlation matrix, with significance values. R's default analysis technique is a Pearson correlation.

In SPSS, the excel file of the data was imported into SPSS. Then the bivariate correlation function was selected, and all the variables were put into the argument box. I then selected Pearson correlation and a two tailed test of significance. Missing values also had to be excluded pair wise rather than list wise. Due to the large dataset and several missing values, losing data by excluding pair wise would keep the data significant.

Multivariate Regression

In R, in order to run a multilevel regression, and to receive detailed information, the lm Support version 2.9.13 package needed to be installed to RStudio. This package is written and maintained by John Curtin. The data was imported into RStudio. Some of the models were comparing ratios. Rather than computing the ratios directly in R, they were computed in the Excel file as a new column, which was used in the code. Objects for each linear model were created using the lm() function in the base package of r. These two models were compared using the model Compare() function from the lm Support package. The function gives SSE, change in R^2 and a p value. The base package of r does not give standardized coefficients in the result. The QuantPsyc. Version 1.5 package had to be installed into RStudio to find the standard coefficients. The package was created and maintained by Thomas D. Fletcher. This allows us to compare how much impact each of the predictor variables has on labor costs.

In SPSS the multilevel multivariate regression is similar to a single level multivariate regression in SPSS.

The excel file of the data had to be imported into the software. The linear regression function is used. The control model is put in the first level of independent variables, and the testing model is put in the second level of independent variables. Missing data are excluded pair wise and R squared changes are noted.

Models

After I ran the correlation matrix, I found that labor costs are strongly correlated with employees, total assets, and total liabilities. Because these were such strongly correlated, they were the control model in models 1 and 2 (see Table 2A). This allowed me to test how significantly the other variables with smaller correlations are related to labor costs. Model 1 is a hierarchical regression analysis using ratios to determine the effect on labor cost. The ratio measure for the organizational level factors is over net sales. Model 1 consists of the control model, current ratio, quick ratio, cash flow margin, degree of operating leverage, debt to assets, working capital turnover, operating income after depreciation/net sale, capital expenditure/net sale, cost of goods sold/net sale, and selling and general administrative costs/net sale. Model 2 consists of absolute values as a comparative technique, rather than ratios. Model 2 is the control model with operating income after depreciation, capital expenditure, and cost of goods sold, selling and general administrative costs, working capital, net sale, cash flow, current assets, and current liabilities.

Another measurement for comparative ratios is total assets, rather than net sale. In Table 2B, Model 1A, and the respective control model are identical to Model 1, with the position of net sale and total assets, switched. Model 1A consists of the control model, current ratio, quick ratio, cash flow margin, degree of operating leverage, debt to assets, working capital turnover, operating income after depreciation/total assets, capital expenditure/total assets, cost of goods sold/total assets, and selling and general administrative costs/total assets.

Exporting Results

The sink() function in the base package of R was used to export results from R. The first command line is sink(), and the desired title of the exported text file is put in as a parameter. Then, everything, which wants to be exported, needs to be put inside of a print function. Finally, another empty sink() function needs to be placed in order to indicate the end of the text file. After being exported into a text file, the data was imported into excel for easier analysis. The import command was used in Excel and the type of file should be a text file. From there, I selected the file I wished to import. I selected that the data is delimited; meaning separated by certain indicators, and selected those indicators to be tabs and spaces. The data is then imported into excel and divided by category for easier analysis and transferring.

Because results in SPSS are exported as tables in PDFs, they do not need to be imported into Excel. First, I selected everything I wanted to export and hit the export button in the toolbar. Then, I chose that I only wanted to export what I selected and changed the directory to my destination. The results are then exported to the computer.

RESULTS

After installing multiple packages into RStudio to expand the statistical techniques used, R gave several histograms, a detailed correlation matrix, multivariable regression and a multilevel regression analysis. SPSS did not need any additional packages. All the functions are built into the software. The histograms (Tables 1A-H) showed that many variables have skewed distribution. The ordinary least squares regression results were similar between original variables and their natural log transformations; so, we used original variables. This would allow for better interpretation of regression coefficients for practitioners.

Correlation Matrix

The Hmisc package in R gave a correlation matrix between all the transformed variables. The bivariate correlate function gave a correlation matrix between all the input variables. It gave a table with the Pearson correlation, significance, and the number of values. Both R and SPSS gave the same values in their matrices. Many of the variables were significantly correlated. There was a 0.95 correlation between Labor costs and employees, which is logical because labor costs increase as more employees are hired. Labor cost also had high positive correlations with assets, liabilities, capital expenditure, sales, cost of goods sold, selling and general administrative expenses, operating income after depreciation and cash flow. Labor cost is negatively related to current ratio and quick ratio. R also gave the significance values for all the relationships. Most of the relationships were significant. Labor costs did not have a

VOL: 4, ISSUE: 12
December/2018
https://ijbassnet.com/
E-ISSN: 2469-6501

significant relationship with working capital turnover and degree of operating leverage.

Multivariate Regression

In R, the lm Support package allowed me to compare various models. Each new model I made was a different level of the regression. The missing values in the data file had to be filled in with the mean because the model compare() function needed the number of values in each model to be the same. The summary() function gave the unstandardized coefficients, with significance values. In SPSS, to keep the analysis comparable, the lm.beta() function gave the standardized coefficients. Both R and SPSS detected identical relationships. Model 1 displayed significant positive relationship between labor cost and employees, assets, quick ratio, and cost of goods sold/net sale. Labor cost is negative associated with total liabilities, current ratio, cash flow margin, debt to assets, operating income after depreciation/net sale, capital expenditure/net sale and selling and general administrative expenses/net sale. The multivariate model 1 explained 69.86% of the variability in the data.

Model 1A demonstrated similar results. However, this time, R and SPSS detected labor cost to be positively related to liabilities, operating income after depreciation/total assets and selling and general administrative expenses/assets total. 71.12% of the variability in the data was explained by model 1A. Model 2 detected a smaller, yet still significant, positive relationship between labor cost and total assets and total liabilities. It also found labor cost to be positively related to cost of goods sold, selling and general administrative expenses, current assets and current liabilities. There was a strong relationship to cash flow. This model explained 76.8% of the variability in the data.

DISCUSSION

These analytics result show that labor cost is influenced by many firm level variables. These provide interesting and useful guidelines to researchers and practitioners on how a firm allocates its resources. For example, labor cost is positively related to assets but negatively related to capital expenditure. This indicates that firms that have higher levels of total assets also spend more on labor. However, firms that are investing heavily in capital equipment, maybe replacing labor with capital. It also shows that high labor cost negatively impacts net sales. This is understandable, as firms spend more on labor, their net revenue earnings from sales declioine. Taken together, these analytics given a broad new picture of labor cost allocation behavior of the firm.

Both R and SPSS had identical outputs in analyzing organizational level factors and labor cost. The analyses completed were correlation matrix and multivariate regression. In the correlation matrix, both R and SPSS gave correlation coefficients and levels of significance. In the multivariate regression, R gave the unstandardized coefficients, standard error, t score, significance value, and the R/R^2 information in the summary() function of the linear model. It also ran an ANOVA analysis using the built in anova() function. R gave the standardized coefficients using the lm.beta() function, from an external package and the R^2 change using the model Compare() function from another external package. Initially, the functions gave the outputs in scientific notation. After changing the output format with a line of code (shown in appendix 1), all results were in decimal notation. In SPSS, I had to select my output specifications when inputting the variables and it came out in the output as multiple tables. SPSS gave the R/R^2 information, standard error, significance value, f value, unstandardized and standardized coefficients with their t-score. SPSS also ran an anova analysis comparing the levels.

When conducting the multilevel regression in R, I used the model compare() function from the lm Support package. For the function to perform its analysis, each model needed to have the same number of data points. The missing values for each variable were replaced with its mean, computed from a basic R analysis. This was also done in SPSS to keep the data comparable. Future study can use multiple imputation method for missing data in R and see if the results are different.

One observation made when running the analysis was regarding the multivariate regression on ratios. The dataset imported had all the organizational level factors. In R, when coding the comparison model, the ratios were typed directly as one factor/factor (ex. OIADP/SALE). To test if the results were accurate, new objects were created as the ratios and put directly into the model (shown in appendix). When running both modes separately and comparing the results, the two methods showed different r^2 values. So, I put the ratios as new objects to ensure that R was not trying to overestimate the model. Later, it was observed that SPSS gave the same results, showing that R was not modeling the relationships correctly when the ratios were coded directly.

Because we were running a simple analysis, running it in SPSS was very fast, because everything could be selected and transferred to the designated box. Other parameters could be selected from different option boxes. However, in R, it took a little more time because each of the organizational level factors had to be put in with code. Different packages and different functions had to be used to run different types of analysis. However, after working with R, there seem to be a huge variety of packages available to be installed into R. In R, data types changes, objects creation and advanced manipulations can all be done within the software. Further study would include comparing the results of R and SPSS with a more advanced analysis. This would help determine if R would be better at analyzing more complicated data because of the various manipulations, which can be done within it.

Grover and Kar (2017) note that big data can be retrieved from various sources. Unstructured data refers to data that doesn't fit neatly into the traditional row and column structure of relational databases. Examples of unstructured data include: emails, videos, audio files, web pages, and social media messages. In today's world of big data, most of the data that is created is unstructured with some estimates of it being more than 95% of all data generated. Unstructured data analytics with machine-learning intelligence allows organizations to successfully use these data for Business Analytics based decision-making. In the future, researchers can investigate unstructured data in the form of strategic directions etc. and see their effect on labor cost.

CONCLUSION

The purpose this paper was to run big data analytics on labor cost data from all available US firms over past thirty years. The analytics found many significant relationships between labor cost and other organization variables. This would provide guidance to managers in resource allocation decision-making. I compared two powerful analytical tools, R and SPSS. R is a machine learning programming language used for statistical analysis and modeling. It is package based and written in lines of code. It uses functions in which parameters are passed to. It provides a large variety of statistical techniques and provides an Open Source route to participation. SPSS offers statistical analysis, graphing techniques; and also open source extensibility. SPSS has a built-in set of functions in which parameters can be adjusted by selecting different options in prompt boxes.

Both R and SPSS gave the same results when analyzing the panel data of firm years and the cross-sectional data of coefficients of variability. They both have many powerful analytical techniques to analyze large data sets. However, in SPSS, the data input needs to be in a specific format for the software to read it correctly. R reads imperfections in the data and can be fixed within the code. With simple language like instructions, I can manipulate the dataset to work with the analysis I want to run. Therefore, R is more applicable to big data because most of the times big data is quite messy.

This research is important for practitioners because it demonstrates that both R and SPSS are powerful tools for analyzing large data sets. The data set contained panel data for 1212 firms, totaling, 24,186 firm years with 21 organizational level factors. Both software were able to run multivariate regressions and get the same results. Managers and researchers can use the software to draw conclusions about their investigation and use it to make strong managerial decisions.

Further research includes looking at different machine learning software and comparing those results to the results received from R and SPSS. One such software is Weka, which is a free, data mining machine learning atmosphere. However, it is important to understand that this is preliminary research, as there is not a lot of literature regarding the comparison of R and SPSS. Further research involving more advanced analysis techniques will help give a better understanding of the differences in R and SPSS.

REFERENCES

Ballester, M., Livnat, J., and Sinha, N. (2002). Labor costs and investments in human capital. *Journal of Accounting, Auditing & Finance*, 17(4), 351-373.

Blinder, A., Ed. 1990. Paying for Productivity.Washington DC, Brookings Institution.

Callahan, C. M., andStuebs, M. (2010). A theoretical and empirical investigation of the impact of labor flexibility on risk and the cost of equity capital.*Journal of Applied Business Research*, 26(5), 45-62.

de Grip, A., and Sieben, I. (2005). The effects of human resource management on small firms' productivity and employees' wages. *Applied Economics*, 37(9), 1047-1054.

Freeland, M. S., Anderson, G.F. and Schendler, C. 1979. National Input Price Index.*Health Care Financing Review*. 1 (1), 37–61.

Grable, J. E., & Lyons, A. C. (2018). An Introduction to Big Data. *Journal of Financial Service Professionals*, *72*(5), 17–20.

Grover, P., & Kar, A. (2017). Big Data Analytics: A Review on Theoretical Contributions and Tools Used in Literature. *Global Journal of Flexible Systems Management*, *18*(3), 203–229.

Kljenak, D. V., Radojko, L., &Jovancevic, D. (2015). Labor costs analysis in the trade market of Serbia. *Management Research and Practice*, 7(3), 59-79.

Jain, P., Gyanchandani, M., &Khare, N. (2016). Big data privacy: A technological perspective and review. *Journal of Big Data*, 3(1), 1-25.

Li, J., Tao, F., Cheng, Y., & Zhao, L. (2015). Big data in product lifecycle management. *The International Journal of Advanced Manufacturing Technology*, 81(1–4), 667–684.

Liu-Thompkins, Y., & Malthouse, E. C. (2017). A Primer on Using Behavioral Data for Testing Theories in Advertising Research. *Journal of Advertising*, *46*(1), 213–225.

Marsland, S. (2011). *Machine learning: an algorithmic perspective*. Chapman and Hall/CRC.

Osborne, Jason (2013). *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do before and after Collecting Your Data*. ThousandOaks, CA: Sage.

Park, Hun Myoung. (2009). Comparing Group Means: T-tests and One-way ANOVA Using STATA, SAS, R, and SPSS. Working Paper. The University Information Technology Services (UITS) Center for Statistical and Mathematical Computing, Indiana University.

Samudhram, A., Shanmugam, B., & Kevin Lock, T. L. (2008). Valuing human resources: An analytical framework. *Journal of Intellectual Capital*, 9(4), 655-667.

van Buuren, S., &Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45*(3), 1 - 67.

Zhang, D., & Jeffrey J.P. Tsai. (2003). Machine learning and software engineering.*Software Quality Journal*, 11(2), 87-119.

FIGURE 1A: LABOR COST

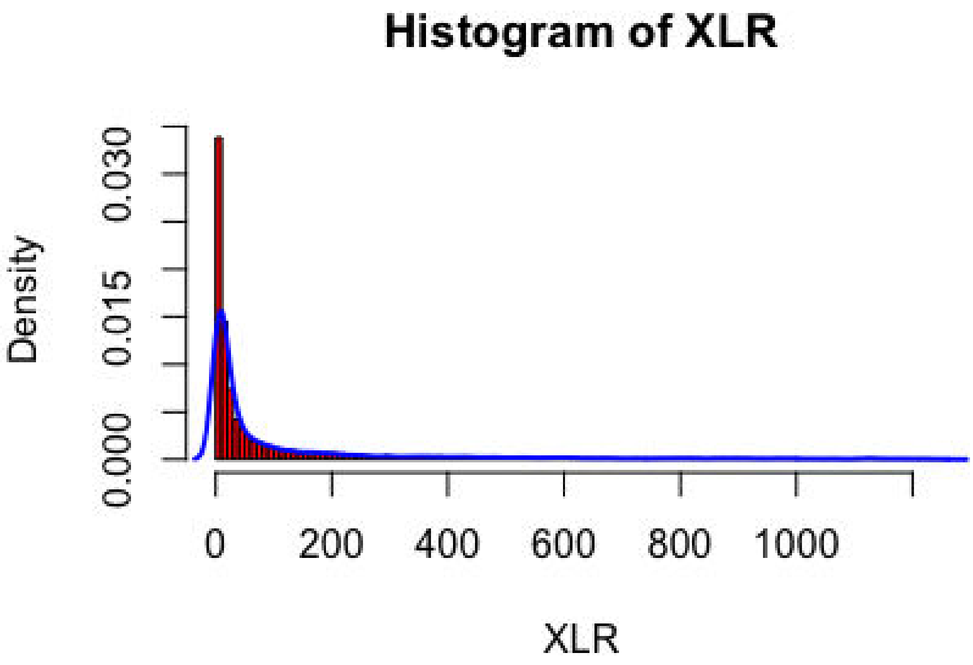

FIGURE 1B: NUMBER OF EMPLOYEES

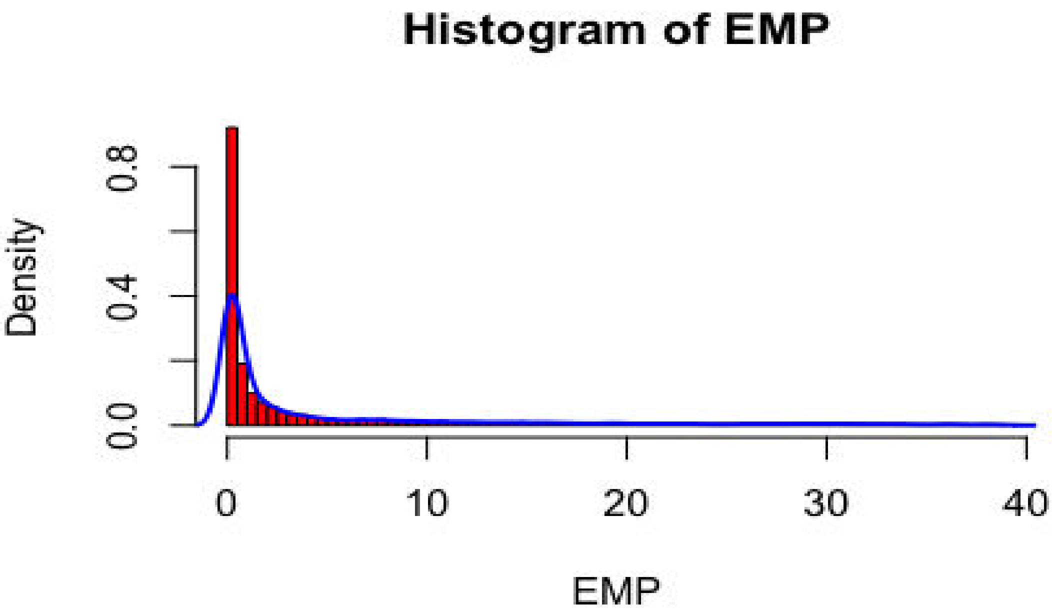

FIGURE 1C: CASH FLOW

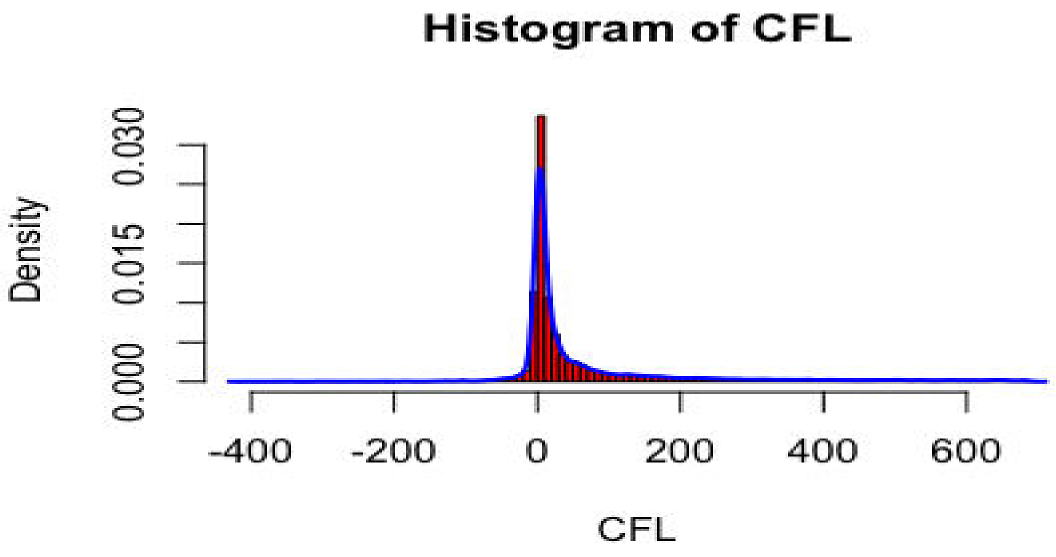

FIGURE 1D LIABILITY TOTAL

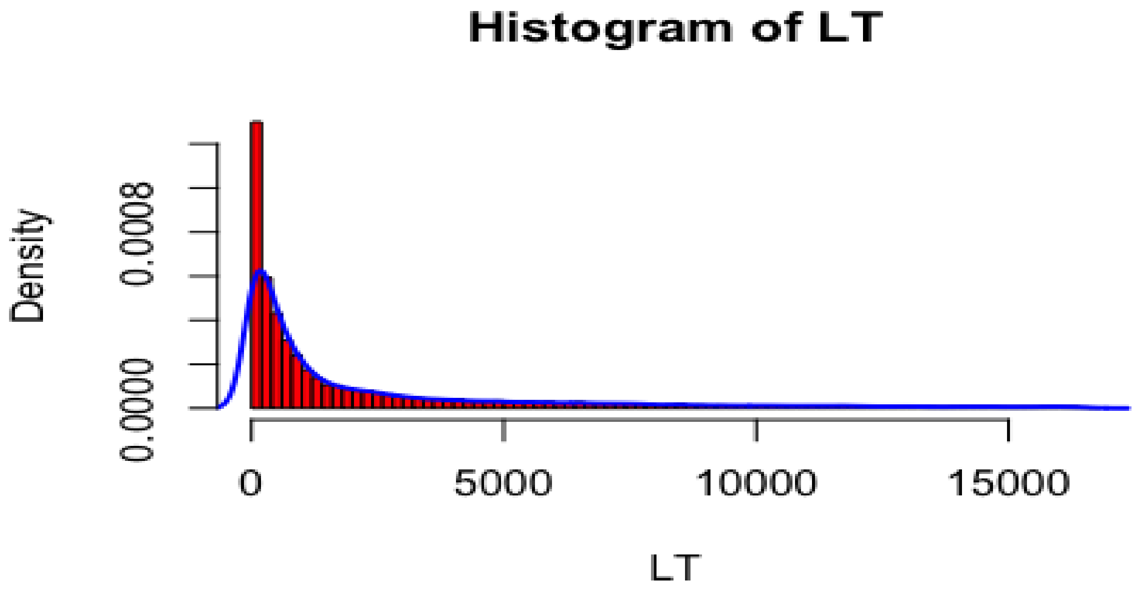

FIGURE 1E: ASSETS TOTAL

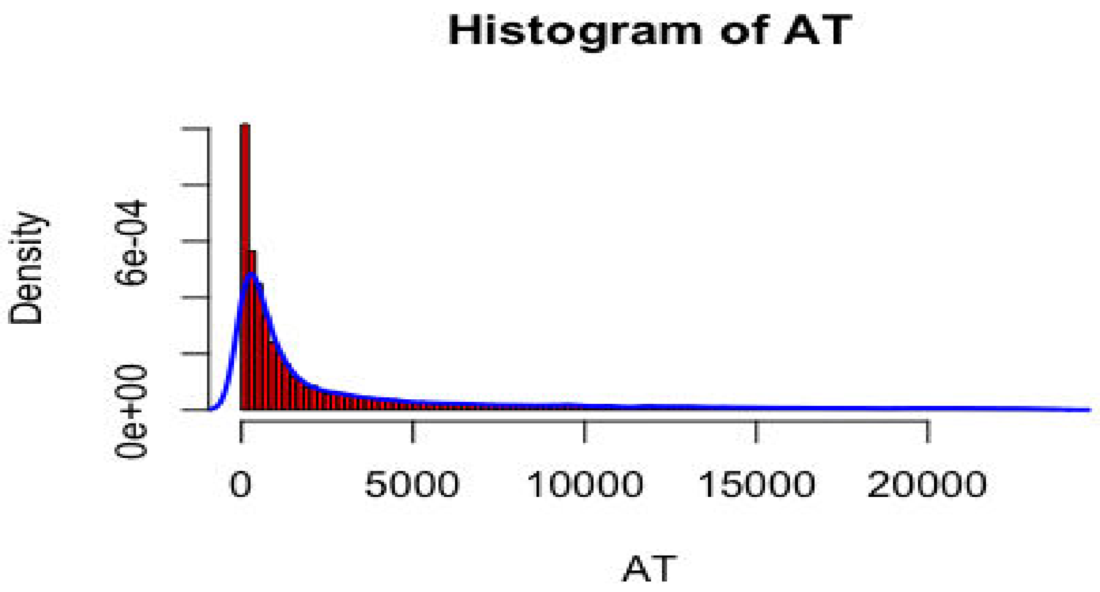

FIGURE 1F: CAPITAL ESPENDITURE

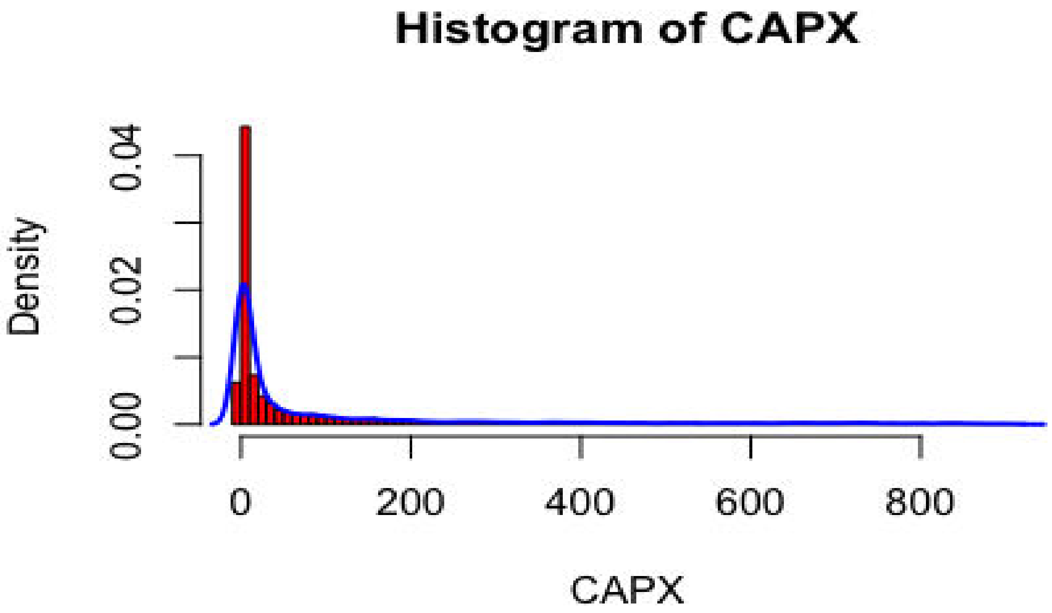

FIGURE 1G: COST OF GOODS SOLD

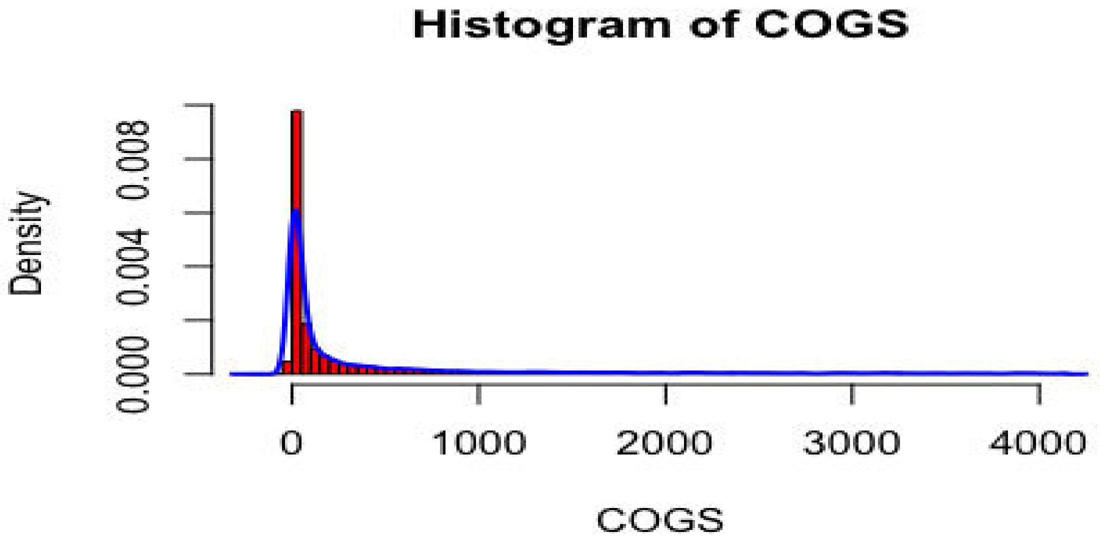

FIGURE 1H: NET SALES

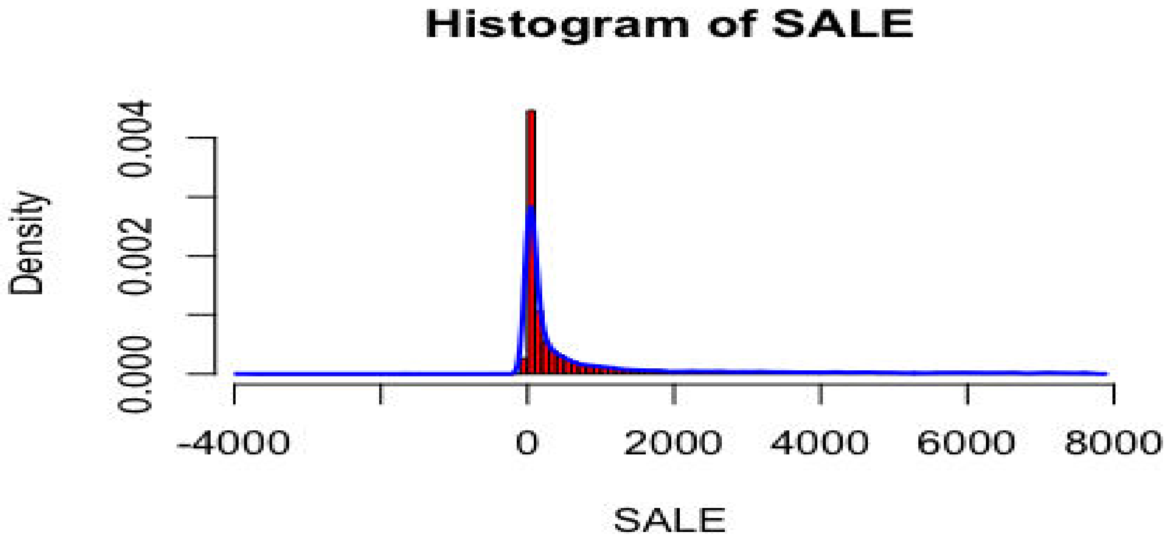

TABLE 1

Descriptive Statistics and Correlation Coefficients for Model 1, 1A, 2

	Variable	Mean	SD	1	2	3	4	5
1.	Labor Cost	4.15	2.68					
2.	Employees	0.49	2.71	0.95				
3.	Assets Total	7.33	2.75	0.84	0.77			
4.	Liabilities Total	6.92	2.96	0.79	0.73	0.97		
5.	Current Assets	5.4	3.23	0.91	0.87	0.96	0.93	
6.	Current Liabilities	5.02	3.33	0.93	0.89	0.95	0.98	0.95
7.	Capital Expenditure	3.25	3.37	-0.86	0.88	0.81	0.76	0.89
8.	Net Sales	5.88	2.79	0.96	0.93	0.87	0.82	0.94
9.	Cost of Goods Sold	5.09	2.97	0.95	0.93	0.79	0.73	0.93
10.	Selling and General Administrative Expenses	3.76	3.53	0.71	0.66	0.63	0.61	0.82
11.	Operating Income After Depreciation	4.49	2.51	0.89	0.82	0.9	0.85	0.92
12.	Cash Flow	3.97	2.65	0.92	0.87	0.86	0.79	0.93
13.	Debt to Assets	-0.43	0.74	0.06	0.05	0.2	0.43	0.14
14.	Current Ratio	0.29	1.02	-0.15	-0.16	-0.04	-0.22	0.07
15.	Quick Ratio	-0.03	1.18	-0.16	-0.18	-0.08	-0.25	0.04
16.	Degree of Operating Leverage	-0.08	0.86	0.22	0.2	0.12	0.09	0.24
17.	Degree of Operating Leverage-Median	0.37	0.48	-0.26	-0.23	-0.18	-0.16	-0.33
18.	Degree of Operating Leverage-Dummy Variable	408.01	2978.73	0.14	0.12	0.16	0.15	0.21
19.	Working Capital	159.74	13576.42	0	0	0	0	0
20.	Working Capital Turnover	-95.69	3772.98	0.08	0.06	0.06	0.06	0.08
21.	Cash Flow Margin	11.89	1045.5	-0.01	-0.01	0	0	0.01

N = 24186, all correlations above .02 are significant at *p* < .05.

Variable	6	7	8	9	10	11	12	13
1. Labor Cost								
2. Employees								
3. Assets Total								
4. Liabilities Total								
5. Current Assets								
6. Current Liabilities								
7. Capital Expenditure	0.9							
8. Net Sales	0.94	0.91						
9. Cost of Goods Sold	0.94	0.89	0.96					
10. Selling and General Administrative Expenses	0.8	0.8	0.68	0.63				
11. Operating Income After Depreciation	0.91	0.84	0.93	0.86	0.63			
12. Cash Flow	0.93	0.91	0.96	0.92	0.76	0.96		
13. Debt to Assets	0.36	0.24	0.04	0.04	0.11	0.02	-0.06	
14. Current Ratio	-0.24	-0.2	-0.11	-0.14	-0.1	-0.21	-0.24	-0.7
15. Quick Ratio	-0.27	-0.21	-0.14	-0.16	-0.16	-0.22	-0.26	-0.7
16. Degree of Operating Leverage	0.22	0.16	0.21	0.21	0.1	0.07	0.11	-0.09
17. Degree of Operating Leverage-Median	-0.31	-0.13	-0.22	-0.19	-0.15	-0.16	-0.18	0.07
18. Degree of Operating Leverage-Dummy Variable	0.15	0.14	0.16	0.15	0.16	0.17	0.18	-0.01
19. Working Capital	0	0	0	0	0	0	-0.01	0.01
20. Working Capital Turnover	0.06	0.06	0.1	0.04	0.04	0.07	-0.01	0.01
21. Cash Flow Margin	0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	0

N = 24,186, all correlations above .02 are significant at $p < .05$

	Variable	14	15	16	17	18	19	20	21
1.	Labor Cost								
2.	Employees								
3.	Assets Total								
4.	Liabilities Total								
5.	Current Assets								
6.	Current Liabilities								
7.	Capital Expenditure								
8.	Net Sales								
9.	Cost of Goods Sold								
10.	Selling and General Administrative Expenses								
11.	Operating Income After Depreciation								
12.	Cash Flow								
13.	Debt to Assets								
14.	Current Ratio								
15.	Quick Ratio	0.95							
16.	Degree of Operating Leverage	0.03	0.02						
17.	Degree of Operating Leverage-Median	0	0	-0.63					
18.	Degree of Operating Leverage-Dummy Variable	0.16	0.13	0.1	-0.09				
19.	Working Capital	0	0	-0.01	0.01	0			
20.	Working Capital Turnover	0.03	0.04	0.09	-0.03	0.01	0		
21.	Cash Flow Margin	0	0	0.01	-0.01	0.01	0	0	

N = 24,186, all correlations above .02 are significant at $p < .05$.

TABLE 2A
Results of Hierarchical Regression Analysis of Labor Cost
(Model 1with ratio measures/Net sale, Model 2 is original variables)

Variable	Control Model	SPSS Model 1	R Model 1	SPSS Model 2	R Model 2
Employees	0.66249415***	.654***	0.6538715805***	.403***	0.402781542***
Assets Total	0.30617752***	0.275***	0.275141234***	.045**	0.045131377**
Liabilities Total	-0.06436573***	-.018	-0.018143308	.052***	0.052280969***
Current Ratio		-.065***	-0.064654774***		
Quick Ratio		.021**	0.021122929**		
Cash Flow Margin		-.016***	-0.01586558***		
Degree of Operating Leverage		-.001	-0.000777039		
Debt to Assets		-.041***	-0.040598076***		
Working Capital Turnover		.001	0.000969402		
Operating Income After Depreciation/Net Sales		.012**	0.012049949**		
Capital Expenditure/Net Sales		-.032***	-0.032357131***		
Cost of Goods Sold/Net Sales		.018***	0.018058063***		
Selling and General Administrative Expenses/Net Sales		-.011**	-0.010740364**		
Operating Income After Depreciation				-.050 ***	-0.049519439***
Capital Expenditure				-.174 ***	-0.174082544***
Cost of Goods Sold				.292 ***	0.291646311***
Selling and General Administrative Expenses				.051 ***	0.050889002***
Working Capital				-.007 *	-0.006945238*

Variable	Control Model	SPSS Model 1	R Model 1	SPSS Model 2	R Model 2
Net Sales				-.036 **	-0.035833435**
Cash Flow				.283 ***	0.282548121***
Current Assets				.071 ***	0.071482121***
Current Liabilities				.048 ***	0.047892101***
Adjusted R^2	0.6958	0.699	0.6986	0.768	0.768
R^2 change		.003 ***	0.002937571***	.072 ***	0.07234069***

Standardized coefficients are reported
Control Model reported from R, but similar in SPSS
*** p < .001 ** p< .01 * p < .05

TABLE 2B
Results of Hierarchical Regression Analysis of Labor Cost (with ratio measures/Total assets)

Variable	Control Model	SPSS Model 1A	R Model 1A
Employees	0.583605***	.580 ***	0.580126263***
Net Sale	0.24732534***	.232 ***	0.23375826***
Liabilities Total	0.08923874***	.112 ***	0.111774911***
Current Ratio		-.057 ***	-0.057872184***
Quick Ratio		.037 ***	0.037146097***
Cash Flow Margin		-.025***	-0.025303533***
Degree of Operating Leverage		-.001	-0.000522175
Debt to Assets		-.033***	-0.033589076***

Variable	Control Model	SPSS Model 1A	R Model 1A
Working Capital Turnover		.001	0.000795611
Operating Income After Depreciation/Assets Total		.013***	0.028404509***
Capital Expenditure/ Assets Total		-.032 ***	-0.01064726*
Cost of Goods Sold/ Assets Total		.003	-0.01332804**
Selling and General Administrative Expenses/ Assets Total		.024 ***	0.034460081***
Adjusted R^2	0.7086	.711	0.7112
R^2 change		**.002*** **	**0.002775482*** **

Standardized coefficients are reported
Control Model reported from R, but similar in SPSS
*** p < .001 ** p< .01 * p < .05

APPENDIX 1
R Code for Pre-analysis setup

Changing Outputs to Decimals

```
options(scipen=999)
```

Installing and loading packages

```
install.packages("QuantPsyc")
install.packages("lmSupport")
install.packages("Hmisc")
library(QuantPsyc)
library(lmSupport)
library(Hmisc)
```

APPENDIX 2
R Code for Panel Data analysis

Importing Data

```
    Data<-read.csv(file.choose(), header=TRUE, sep=",", stringsAsFactors = FALSE)
    attach(Data)
    names(Data)
    class(XLR)
```

Correlation Matrix

```
    Data<-as.matrix(Data)
    rcorr(Data)
```

Multivariate Regression

Model 1

```
        OIADP.SALE<-OIADP/SALE
        CAPX.SALE<-CAPX/SALE
        COGS.SALE<-COGS/SALE
        XSGA.SALE<-XSGA/SALE
        controlModel1<-lm(XLR~EMP+AT+LT)
        model1<-
        lm(XLR~EMP+AT+LT+OIADP.SALE+CR+QR+CFLM+DOL+DebtToAssets+CAPX.SALE+COG
        S.SALE+XSGA.SALE+WCTO)
        summary(controlModel1)
        summary(model1)
        lm.beta(controlModel1)
        lm.beta(model1)
        modelCompare(controlModel1, model1)
        anova(controlModel1, model1)
```

Model 1A

```
        OIADP.AT<-OIADP/AT
        CAPX.AT<-CAPX/AT
        COGS.AT<-COGS/AT
        XSGA.AT<-XSGA/AT
        controlModel1A<-lm(XLR~EMP+SALE+LT)
        model1A<-
        lm(XLR~EMP+SALE+LT+(OIADP.AT)+CR+QR+CFLM+DOL+DebtToAssets+(CAPX.AT)+(COG
        S.AT)+(XSGA.AT)+WCTO)
        summary(controlModel1A)
        summary(model1A)
        lm.beta(controlModel1A)
        lm.beta(model1A)
        modelCompare(controlModel1A, model1A)
        anova(controlModel1A, model1A)
```

Model 2

```
controlModel2<-lm(XLR~EMP+AT+LT)
model2<-lm(XLR~EMP+AT+LT+OIADP+CAPX+COGS+XSGA+WC+SALE+CFL+CA+CL)
summary(controlModel2)
summary(model2)
lm.beta(controlModel2)
lm.beta(model2)
modelCompare(controlModel2, model2)
anova(controlModel2, model2)
```

Calculating quartiles

```
quantile(XLR, na.rm-TRUE)
quantile(EMP, na.rm-TRUE)
quantile(AT, na.rm-TRUE)
quantile(LT, na.rm-TRUE)
quantile(CAPX, na.rm-TRUE)
quantile(SALE, na.rm-TRUE)
quantile(COGS, na.rm-TRUE)
quantile(CFL, na.rm-TRUE)
```

Plotting Data

```
hist(XLR, breaks=100, col="red", prob=TRUE)
lines(density(XLR, na.rm = TRUE), col="blue", lwd=2)

hist(EMP, breaks=100, col="red", prob=TRUE)
lines(density(EMP, na.rm = TRUE), col="blue", lwd=2)

hist(AT, breaks=100, col="red", prob=TRUE)
lines(density(AT, na.rm = TRUE), col="blue", lwd=2)

hist(LT, breaks=100, col="red", prob=TRUE)
lines(density(LT, na.rm = TRUE), col="blue", lwd=2)

hist(CAPX, breaks=100, col="red", prob=TRUE)
lines(density(CAPX, na.rm = TRUE), col="blue", lwd=2)

hist(SALE, breaks=100, col="red", prob=TRUE)
lines(density(SALE, na.rm = TRUE), col="blue", lwd=2)

hist(COGS, breaks=100, col="red", prob=TRUE)
lines(density(COGS, na.rm = TRUE), col="blue", lwd=2)

hist(CFL, breaks=100, col="red", prob=TRUE)
lines(density(CFL, na.rm = TRUE), col="blue", lwd=2)
```

VOL: 4, ISSUE: 12
December/2018
https://ijbassnet.com/
E-ISSN: 2469-6501

APPENDIX 3
Importing Data into R

R has many functions that allow one to import data from other applications. The following table lists some of the useful text import functions, what they do, and examples of how to use them.

Function	What It Does	Example
read.table()	Reads any tabular data where the columns are separated (for example by commas or tabs). You can specify the separator (for example, commas or tabs), as well as other arguments to precisely describe your data.	read.table(file=”myfile”, sep=”t”, header=TRUE)
read.csv()	A simplified version of read.table() with all the arguments preset to read CSV files, like Microsoft Excel spreadsheets.	read.csv(file=”myfile”)
read.csv2()	A version of read.csv() configured for data with a comma as the decimal point and a semicolon as the field separator.	read.csv2(file=”myfile”, header=TRUE)
read.delim()	Useful for reading delimited files, with tabs as the default separator.	read.delim(file=”myfile”, header=TRUE)
scan()	Allows you finer control over the read process when your data isn’t tabular.	scan(“myfile”, skip = 1, nmax=100)
readLines()	Reads text from a text file one line at a time.	readLines(“myfile”)
read.fwf	Read a file with dates in fixed-width format. In other words, each column in the data has a fixed number of characters.	read.fwf(“myfile”, widths=c(1,2,3)

In addition to these options to read text data, the package foreign allows you to read data from other popular statistical formats, such as SPSS. To use these functions, you first have to load the built-in foreign package, with the following command:

```
> library("foreign")
```

The following table lists the functions to import data from SPSS, Stata, and SAS.

Function	What It Does	Example
read.spss	Reads SPSS data file	read.spss(“myfile”)
read.dta	Reads Stata binary file	read.dta(“myfile”)
read.xport	Reads SAS export file	read.export(“myfile”)

Source:
https://www.dummies.com/programming/r/the-benefits-of-using-r/