**CSCI.4923 Capstone in Interprofessional Informatics**

**Analysis and Conclusion**

Sachin Karki

Texas Woman’s University

Denton, Texas

Author Note

Sachin Karki, Texas Woman’s University, Mathematics & Computer Science

**Introduction**

The crucial part of any research or study is to explore evidence to interpret findings and reach conclusions. Research involving statistical evidence needs analysis for sure. In this case, testing and analyzing the dataset is necessary to extract and retrieve any information that helps in answering our research question. Our datasets, as mentioned and described previously *Effects.xlsx, Activity.xlsx and Scores.xlsx* will be analyzed extensively with different tools and procedures to test our hypothesis and check the association with other related variables. The outcome of the analysis will help in answering the following research questions.

1. What interpersonal behaviors candidates practiced that were susceptible to excessive smartphone usage?
2. Is there any relationship between different personality types and excessive smartphone usage?
3. What activities of smartphones contributed to excessive usage in different personality types?
4. Which personality type is more vulnerable to the effects of excessive smartphone usage?

This paper is an in-depth exploration and interpretation of the analysis and its results.

**Methods Used**

The originally retrieved dataset was huge with its content so it was realized immediately that the datasets should be broken down to ease the analysis. After new datasets were created with relevant sets of variables, methods of analysis were explored for each dataset. The new datasets were also heavy with its content, so it was certain that each dataset had to go through multi-layered approach of analysis to reach the conclusion. Each layer or step of analysis determined a unique insight which was used later with other variables to test different outcomes likes variable dependency, association, correlation, and regression.

Here’s the breakdown of various forms of analysis.

1. Variable dependency check with Correlation Matrix
2. Variable sorting with Cohen’s D
3. Variable selection with Akaike Information Criterion (AIC) and Error rate reduction with Backward Algorithm
4. Association between dependent and independent variables with Multinomial and Linear Regression

**Description and Interpretation of Analysis**

As mentioned previously, the analyses were based on statistical and visual findings. Using the three datasets that were created and curated from the original datasets, various methods mentioned earlier were used. Here’s a breakdown of what tools, procedures, and methods were used for both forms of analysis.

1. **Statistical Analysis**

The solid foundation of insights came from statistical analysis which was done mostly with the R programming language in RStudio. SAS was also utilized initially as a reference, but R accumulated better results, so it was preferred to use effectively and extensively. The first two datasets *activity.xlsx and effects.xlsx* were utilized to draw a series of conclusions. To begin with, the analysis was first done in the *activity.xlsx* datasets and the same procedures and methods were applied in the *effects.xlsx* dataset. Here’s the exploratory breakdown of those analyses.

1. ***Activity.xlsx***

First and foremost a correlation matrix was generated to check the dependency between the variables. It is only after we know for sure if a relationship exists between the variables, we can proceed further. At this point the dataset has not been filtered out for any missing data in the dataset so the empty values were treated as the default “Not Available (NA)” by default R procedure to generate a correlation matrix.

***Fig 1: Correlation matrix of activity.xlsx dataset***

The given plot above is the summary of correlation coefficients of a different set of variables. It can be noticed that variables “CallFamMem” and “CallFrnds” have a strong positive relationship. The visible and noticeable intensity of the blue dot tells us that they share strong positive relations and logically speaking they attribute somewhat similar activity of smartphone use. Calling friends and calling family members can be considered as a single variable of just calling in general. Therefore, no analysis is needed for this variable since we found relation among them. The rest of the variables however display a curious pattern so we will proceed with a few more steps to extract conclusions. The code for the correlation matrix is depicted below.

This can be further elaborated by the Spearman coefficient score depicted below.

***Fig 2: Correlation graph of activity.xlsx dataset***

From the correlation graph above, we can notice that variables “CallFamMem” and “CallFrnds” have a Spearman score of 0.73 which represents an above-average number on a scale of 1 so statistically speaking these two variables do not have to be included further for more testing.

This was generated using a basic *tidyverse* package in R and the code is depicted below.

In addition to this, our dataset has missing data as well. The number of participants choosing not to respond on a certain topic is very prominent and as a result, we have variables with simply no input. Missing data should always be statistically identified and filtered out to avoid high bias in the results. Using the *ggplot* package in R, missing values were identified, and here’s a visual look at the number of missing data.

***Fig 3: Bar graph of missing data in activity.xlsx datasets***

From the graph above it can be verified that the variables like “Rlgn\_Prctces” and “Study\_Purposes” have the highest count of missing data. Thus, they will be excluded from the analysis.

The second step of the analysis is to sort the variables as our dataset consists of categorical values. The method of Cohen’s D was to do so. The code and result are provided below.

***Fig 5: Summary of standard mean difference with Cohen’s D by personality type***

Cohen’s D helps measure the size of the effect of variables and accompanies additional tests like t-test and ANOVA. A value of 0.2 is considered a small effect size in Cohen’s D and from our result above the variable “smoking” has a standard mean difference of 0.2 in personality type.

The third step of our analysis is variable selection. To do so, AIC with a backward algorithm was used to select variables and reduce the error rate. AIC is an estimator of prediction error. The lower the AIC model is, the better it is for our analysis. First variable selection was done with variables like “ID” and “CallFrnds” since the “CallFrnds” retained the highest correlation of 0.73 to the variable “CallFamMem”. This secures the independence of variables. Additionally, “Smoking”, “AGE”, “Study\_Purposes” were selected with a reduced AIC of 503 from the original value of 521.69. Like mentioned earlier, the lower the better. Here’s a look at the procedure and result.

***Fig 6: AIC with a backward algorithm***

The final model above suggests that a one-point increment in age will give us a 0.88 odd ratio (OR) increase in both personality types. Smoking is the strongest factor with an odds ratio of 1.78. Each result was accompanied by a p-value of 0.05 significance.

Furthermore, multinomial regression was used on the variable “ExcessveSmrtPhn\_Use” as it has four levels of response denoting the general admittance of excessive smartphone usage by the participants. Here’s the index of the scale used for four levels of response.

1 = strongly disagree, 2 = somewhat disagree, 3 = somewhat agree, 4 = strongly agree,

For this regression variables, “ID”, “Study\_Purposes”, and “Rlgn\_Prctces”, are excluded, because of the high missing value rate to avoid the biased result. Response level 2 will be used as a reference to compare with other response levels since 2 is a middle ground between participants who agree and disagree with their excessive smartphone usage. The code and the result of the procedures are depicted below.

***Fig 7: Multinomial regression on response levels 1 and 3 at 0.05 significance level***

***Fig 8: Multinomial regression on response level 4 at 0.05 significance level***

Presented results above are done at a significance level p < 0.05. The collective summary of our multinomial regression on all three-response levels (i.e. 1,3.4) with respect to response level 2 indicates the following insights.

* The relative risk ratio for one-unit increase in the variable “Txtng” is 1.35 for being in ExcessveSmrtphn\_Use:1 vs ExcessveSmartPhn\_Use2.
* The relative risk ratio for one-unit increase in the variable “Entertainment” is 1.88 for being in ExcessveSmrtphn\_Use:3 vs ExcessveSmartPhn\_Use2.
* The relative risk-ratio for one-unit increase in the variable “Txtng” is 2.50 for being in ExcessveSmrtphn\_Use:3 vs ExcessveSmartPhn\_Use2.

Lastly, a generalized linear regression was conducted on “ExcessveSmrtPhn\_Use” in respect with other variables like “Gender”, “Smoking”, “CallFamMem”, “Entertainment”, and “OthRsns” at the same four-level of response used for multinomial regression. Here’s a breakdown of code and results

***Fig 9: Linear Regression on all variables from the activity.xlsx datasets***

***Fig 10: Linear regression on selected variables on three response level***

The depicted analysis above was done at a significance level of 0.05. To summarize, here are the key insights from them.

* One-unit increase in the variable “OthRsns” is associated with the decrease in the log-odds of being in ExcessveSmrtPhn\_Use:1 vs. ExcessveSmrtPhn\_Use:2 in the amount of -0.40
* One-unit increase in the variable “Entertainment” is associated with the increase in the log-odds of being in ExcessveSmrtPhn\_Use:3 vs. ExcessveSmrtPhn\_Use:2 in the amount of 0.62
* One-unit increase in the variable “Entertainment” and “OthRsns” is associated with the increase in the log-odds of being in ExcessveSmrtPhn\_Use:4 vs. ExcessveSmrtPhn\_Use:2 in the amount of 0.72 and 0.71 respectively.

Based on the analysis and the third insight listed above, it seems that participants who strongly agree with excessive smartphone usage have contributing factors such as entertainment and other reasons.

1. ***Effects.xlsx***

The second dataset *effects.xlsx* was treated pretty much the same way as the *activity.xlsx* dataset with few exclusions to some of the steps since this dataset showed some prominent results in the preliminary process. To begin with, a correlation matrix was generated to check the variable dependency. It was found that variables like “TotAddiction\_Score”, “Compulsive\_Behaivor”, “Depression\_Score”, “Anxiety\_score” and “Functional\_Impairment” correlated with high scores. Here’s a look at the code and correlation matrix and correlation graph and that was used in R.

***Fig 11: Correlation matrix of effects.xlsx dataset***

From the correlation matrix above, it can be noticed that variables like “TotAddiction\_Score”, “Compulsive\_Behaivor”, “Depression\_Score”, “Anxiety\_score” and “Functional\_Impairment” show a high correlation. The dependency test on these variables can be skipped for further variable selection. This data can also be illustrated with the Spearman correlation score below.

***Fig 12: Correlation graph of effects.xlsx dataset***

Similarly, missing data from the dataset were identified using the *ggplot* package within R. The bar showing the missing value is depicted below.

***Fig 13: Bar graph of missing values from effects.xlsx dataset***

The next step was to clean and sort the data. To do so, Cohen’s D was used one more time. The code, results, and outcome are given below.

***Fig 14: Summary of standard mean difference with Cohen’s D by personality type in effects.xlsx dataset***

From the collective results generated by the Cohen’s D procedure above, the variables

“DecreasdTmeSlpQulty\_SmrtPhneUse”, “Compulsive\_Behavior” and “Functional\_Impairment” retained a perfect size effect of 0.2.

Furthermore, AIC with a backward algorithm was used for variable selection. Choosing a lower model would help in selecting variables that allow us to use ANOVA and Chi-Square for our categorical values. The code and the outcome are down below.

***Fig 15: AIC with a backward algorithm in effects.xlsx dataset***

In the AIC analysis done above, variables like “Depression\_score”, “TotAddiction\_Score”, “Compulsive\_Behavior”, “Functional\_Impairment” are excluded since their association with the other variables are larger than 0.7. For this process, variables like “AGE”, “CannotHveMeal\_NosmrtPhn”, “Lttl\_IntrstDoingThngs”, and “NotAble\_Stpworry" are selected after the backward algorithm. The AIC was reduced from 340.6 to 309.3 with a p-value < 0.05 and few observations were made.

* 1.69 OR with “CannotHveMeal\_NosmrtPhn” by personality 0 to 1
* NotAble\_Stpworry shows notable OR between its level from 2 (OR: 5.3) to Personality type
* Model is significant that in-between final variables, there are relevant predictors to variable “Prsnlty\_type”

To summarize that statistical perspective into a visual perspective, here’s a resulting violin plot that was generated in R.

***Fig 16: Violin plot of selected variables in different personality types***

1. **Visual Analysis**

The second form of analysis was done using the visualization software Tableau. For this part, the third dataset *scores.xlsx* was used. Although not many insights were generated with this dataset the ones that were extracted with the process brought crucial, precise, and to-the-point observation related to our research questions. Scores dataset consists of variables relating to different symptoms that participants admitted to having.

Variables like “Compulsive\_Behavior”, “Functional\_Impairment”, “Feel\_Depressed” are used in this dataset. In the original dataset and its analysis, these variables were used to denote the Smart Phone Addiction Indicator (SPAI) scale that was derived using the mean difference and Chi-square of these categorical values. The new visualization that was generated using the new dataset is a mere visual portrayal and depiction of the SPAI with a minor modification in the values. The personality type 0 was changed to A and 1 was changed to B to make it easier to comprehend and explain the results. No other changes were made.

In Tableau, the *scores.xlsx* was located, the first and foremost challenge in extracting visuals was to correctly identify the dataset and its attributes. Categorical values like our dataset always need a few modifications to use it properly and effectively. For an instance, even though our dataset has numerical values that denote the count of a number they are not properly configured to be used as the whole number so the data type needs to be changed. In our case, the dataset was changed to a numeric whole number and a few variables were converted to dimension in order to achieve the right and appealing set of bar graphs. From this dataset, we are trying to find effects and the difference in vulnerabilities among two personality types. Bar graphs and a dashboard were generated, to sum up, the visualization. Results are depicted below.

***Fig 18: Bar graph of the variable “Depression\_score” on two personality types.***

In the depicted graph above, we can see that personality type B has a high score on the variable “Depression \_score” from the SPAI scale. This hints that personality type A is less susceptible to the effects of excessive smartphone usage.

***Fig 19: Bar graph of “Anxiety\_score” on two personality types***

In the depicted bar graph above personality type B has the count of 445 in “Anxiety\_score” from the SPAI scale. Personality type A following up close but nowhere near to personality B. This makes personality type A less susceptible to the effects of excessive smartphone usage. Meanwhile, personality type B is more susceptible to effects of the excessive smartphone use.

***Fig 20: Bar graph of “Compulsive\_Behavior” among two personality types***

From the illustrated bar graph above, it can be seen that personality type B has a whopping 5792 score from the SPAI scale. Personality type A has a score of 3500. Statistically speaking personality type B is 60% more likely to be affected with compulsive behavior as a result of excessive smartphone usage.

***Fig 21: Bar graph of “Functional\_impairment” among two personality types***

Again, personality type B has the highest score of 5350 on the SPAI scale. Personality type A sums up to around 3100 making it about 57% less likely to be affected with functional impairment as a result of excessive smartphone usage.

The dashboard of collective snapshots of these visuals is also illustrated on the next page for easy reference.

***Fig 22: Dashboard of all variables from the scores.xlsx dataset***

The dashboard above provides a collective visual summary from the *scores.xlsx* dataset

**Meaning, Association and Summary of Analysis with Research Questions**

From our multi-layered analysis of the datasets, a lot of results were produced. Probably more than what was initially expected. Inspection and filtration were done to extract only the meaningful, accurate, and relevant results. Not every method that was utilized returned expected results. To begin with, we had three major variables that represented the interpersonal behaviors of the candidates i.e. “Rlgn\_Prctces”, “Smoking” and “Alcohol\_drnk” from the *activity.xlsx* dataset. Early on “Rlgn\_Prctces” was taken out from the further analysis since it had the highest count of missing data so technically the chances of that variable helping in answering one of our research questions was very minimal. When proceeded further with Cohen’s D only “Smoking” had a perfect size effect of 0.2 and “Alcohol\_drnk” retained a lower value of 0.021. Furthermore, “Smoking” returned a higher odd ratio in AIC and consistently better in multinomial regression with 0.65, 1.25, and 1.97 in all three-response levels. The increasing odds ratio tells us that the slightest change in the “Smoking” variable is likely to cause a shift in the overall analysis. Therefore, it can be concluded that among the three variables of interpersonal behaviors, Smoking highly contributes to excessive smartphone usage. Drinking alcohol is less likely to have a higher effect on excessive smartphone usage although it does contribute in a minor way. Religious practices have almost zero or insignificant effect in causing excessive smartphone use.

Similarly, from the second Cohen’s D on the *effects.xlsx* dataset, it was found that variables like “DecreasdTmeSlpQulty\_SmrtPhneUse”, “Compulsive\_Behavior” and “Functional\_Impairment” retained a perfect size effect of 0.2 out of all other variables. It can be summed up that the effects of excessive smartphones like a decrease in quality sleep time, compulsive behavior, and functional impairment are higher than any other mentioned effects. The SD of personality type 1 or B in three variables was also found to be more causing higher variance. Following up close on the similar pattern are variables like “CannotHveMeal\_NosmrtPhn”, “Lttl\_IntrstDoingThngs” that showed a promising trend. “NotAble\_Stpworry” returned a higher size effect value so to further analyze these variables AIC procedure was conducted. Although no significant insights were found with the procedure but generating a violin plot (Fig 16) tells us that personality type 1 or B are likely to have effects like not being able to have a meal without a smartphone and inability to stop worrying. These variables compared to the ones from Cohen’s D can be weak since they initially did not show a perfect size effect, but they do have minimal to broad effect over excessive smartphone use.

Furthermore, the multinomial and linear regression on the *activity.xlsx* dataset on the basis of three response levels generated some more conclusions on the activities of the smartphone. Variables like “Txtng” “OthRsns” and “Entertainment” showed a consistently increasing pattern on each response level (Fig 7,8,9,10) with “Txtng” being the most prominent variable. In other words, texting was the activity that was strongly associated with excessive smartphone usage, entertainment and other reasons follows up on that as they showed a definite and fair association with excessive smartphone usage. Additionally, from the Cohen’s D on the *activity.xlsx* dataset, personality type 1 or B generated higher Standard Mean Difference than type 0 or A which in general can be said that type 1 just has a higher effect on the activities than type 0. Lastly, the visualizations generated from the *scores.xlsx* dataset in Tableau are obvious as the illustrations are very precise and simple. The SPAI score of different variables showed how the number compare against each other in different personality types. Illustrations depicted above, figure 18,19,20,21,22 summarize that personality type 1 or B is likely to score more on the SPAI scale of different effects of excessive smartphone usage. This tells us that personality type 1 or B is more vulnerable to the effects of excessive smartphone usage.

**Limitation and Potential of the Analysis**

As it’s known, the original dataset was raw when it was retrieved initially so the possibilities with the original dataset are endless. Our analysis is also an instance of one of those possibilities since a new dataset was created by splitting required variables. This not only helped to conduct the required analysis but also to reach conclusions and answer the research questions. However, even our study and the analysis have a few limitations that were not addressed or explored.

For an instance, the variable “smoking” was found to be strongly linear to excessive smartphone usage. This is to say that group of people who admitted smoking as an interpersonal habit are prone to be addicted to excessive smartphone use. However, it is not clear what category of smokers (i.e. heavy, light, or chain smokers) are more likely to be excessive smartphone users. This is by default one of the limitations of the dataset and analysis we performed.

Another instance of such limitation is that our dataset does not have any information or explanation on some of the effects that participants had because of excessive smartphone use. For example, the original dataset and the one we created contains variables like “compulsive behavior”, “depression”, “anxiety”, “functional impairment” etc. but the values for these variables were generated based on the candidate's admission or denial on the respective topic. A classic yes or no type of response could be ineffective for a broad subject that those variables are representing. We are not given or told the contributing factors of those variables and it can be biased since any candidate on a given day may give a different response on a specific topic depending on their mood.

In addition to this, other forms of analysis can be done with our dataset to answer more questions. For an instance, a possible question can be related to finding a connection between texting and anxiety in both personality types, finding if there exists any relationship between participants who are depressed see using a smartphone as entertainment purpose as a cause of depression, finding the relationship between compulsive behavior and the number of people who admit smoking can lead to excessive smartphone usage and then comparing it against with people who deny, etc.

Using two or more multi-layered approaches of analysis can help in finding the answers to the questions mentioned above. New dataset creation with relevant variables can be a starting point.

**References**

Boumosleh, J. M., & Jaalouk, D. (2017). Depression, anxiety, and smartphone addiction in university students- A cross sectional study. PLOS ONE 12(8): e0182239. <https://doi.org/10.1371/journal.pone.0182239>

Boumosleh, J. M., & Jaalouk, D. (2017). *Raw data for all participants*. (Version 1). PLOS ONE. <https://doi.org/10.1371/journal.pone.0182239.s001>

Arumugam, N., Selvanayagam, S., & Sathiyasenan, S. T. (2020). The Effects of Smartphone Usage on University Students. *International Journal of Academic Research in Progressive Education and Development*, 9(3), 170–183. : <http://dx.doi.org/10.6007/IJARPED/v9-i3/7960>