**MIS581 Capstone – Business Intelligence and Data Analytics: Module 8: Portfolio Project: Option #1: Capstone Project – Final Report and Slide Presentation: U.S. Organization**

Charles Seglem

Colorado State University Global

21SB-MIS581-1: Capstone – Business Intelligence and Data Analytics

Dr. Orenthio Goodwin

June 6, 2021

**ABSTRACT**

Diet and nutrition are well studied components in disease treatment, prevention, and athletic performance. However, comparatively limited research has been performed in the United States on the role that diet and nutrition plays in mental health. Previous research on the topic has indicated that a link may exist. To further explore this link, this study analyzed the single day nutritional intake, biochemical lab-work, and mental health screener results of over 4400 participants from the United States National Center for Health Statistics’ 2017-2018 National Health and Nutrition Examination Survey to see if the survey results supported the existence of a link between diet and mental health status. Results from several ANOVAs and linear correlations performed during this study did indicate that a significant link may exist. However, predictive model performance was poor and further indicates the need for additional study.

**MIS581 Capstone – Business Intelligence and Data Analytics: Module 8: Portfolio Project: Option #1: Capstone Project – Final Report and Slide Presentation: U.S. Organization**

This paper is the final Portfolio Project report for the MIS581 Capstone – Business Intelligence and Data Analytics course. There are many components that went into the generation of this report. In short, this report includes the following: information about the dataset and organization the dataset was pulled from along with reasoning for the dataset selection; the project’s research null and alternative hypotheses; the tools and techniques used to analyze the data; security, privacy and ethical concerns for using and analyzing the dataset; the analysis results and which of the project hypotheses was supported; the analysis programming code along with screenshots of the analysis outcomes; uploading the project and code to GitHub; and any other components submitted during the Module 5 Critical Thinking Assignment.

**Introduction**

Diet is a well understood component of overall physical health. It influences athletic performance, body weight, and plays a significant role in both disease prevention and treatment. For example, Scurvy is a historically classic nutrient-related disease in which a Vitamin C deficiency results in the breakdown of connective tissue; and if not treated promptly, will result in death. Given the life-threatening but easily preventable nature of Scurvy, many societies have taken steps to ensure such diseases are prevented. For instance, prevention efforts have led to many foods being fortified, like cereal, so as to help populations meet the minimal micronutrient intakes required to ensure this prevention (Fulgoni & Buckley, 2015, p. 3949). However, despite the very clear role diet plays in disease treatment or prevention, comparatively limited research has been done on the role diet plays in mental disorders. It stands to reason that given the significant role diet can play in overall physical health, that diet would also play a critical role in optimum brain function and therefore have the capacity to impact emotional and psychological health. Thus, this project looks to explore exactly that relationship.

**Study Objective**

In short, the research conducted to produce this paper takes selected datasets from the 2017-2018 National Health and Nutrition Examination Survey (NHANES), produced by the National Center for Health Statistics (NCHS), and sees if a significant relationship exists between incidences of depression and certain nutritional/biochemical markers.

**Study Overview**

To study the project objective, this study analyzed a manually combined dataset from the available 2017-2018 NHANES datasets. The combined dataset included a complete record for 4417 participants with a total of 79 unique variables. These unique variables included a participant identifier, results of a mental health screener questionnaire, results of biochemical lab work, and the results of a nutritional intake interview. Then a specific mental health status variable, labeled DPQ020 which asks directly about and rates incidences and severity of depression, was used as the dependent variable with which to study the objective. Details about these variables and the NHANES datasets can be found in the following sections of this paper.

**About the National Health and Nutrition Examination Survey & National Center for Health Statistics**

The National Center for Health Statistics (NCHS) is located in Maryland and is a part of the United States Centers for Disease Control & Prevention and the Department of Health & Human Services. Originally the NCHS was two separate departments, the National Office of Vital Statistics and the National Health Survey, but in 1960 these two departments merged to form the NCHS as it stands today (Centers for Disease Control and Prevention, March 30, 2020, para. 4). The NCHS is well-funded, receiving approximately $160,397,000 from the Federal Government for the 2020 fiscal year (Centers for Disease Control and Prevention, October 6, 2020, para. 1). Anecdotal reports suggest that the NCHS employs between 700 – 750 employees (Wikimedia Foundation, n.d., overview); however, this could not be verified for accuracy.

Ultimately, the mission of the NCHS “…is to provide statistical information that will guide actions and policies to improve the health of the American people” (Centers for Disease Control and Prevention, n.d., para 1). To this end they offer a number statistical services including data collection surveys, research opportunities, and historical documentation on health trends in the United States.

**Author’s Reasons for Selecting the Dataset**

The author selected the NCHS because it offered a number of robust health related datasets. The author has a keen interest in healthcare as his undergraduate degree focused on physiology and biochemistry. The author also hopes to shift his career into data analytics specifically pertaining to data derived from healthcare related research. So he wanted to use this project as an opportunity to gain additional familiarity with performing data analysis on such a dataset. In addition to the public data offered by the NCHS and the author’s personal/career interests in the health sciences, the author selected this organization because the offered data can be used to assess the overall health of people within the United States. Data like this is especially relevant with the COVID-19 pandemic still underway.

Of particular interest to the author is nutrition as it relates to the overall health of an individual and to the prevention and/or treatment of disease. It is well studied and generally known that diet influences incidences of cardiovascular disease (Li et. al., 2020) and can help prevent premature death (Kim et. al., 2018). The Journal of American Medical Association also noted that poor diet is among the leading causes of death and disease burden in the United States (Murray, 2018, p. 1444). Outside of even immediate health concerns, dietary preferences are also sources of significant environmental impact (Auestad & Fulgoni, 2015, p. 19). Therefore, studies on nutrition and diet are among some of the most critical points for healthcare-related research foci. Thus with the author’s previously noted health sciences background, career aspirations, interest in nutrition, and the high level of importance studies relating diet to health has for the United States population as a whole, the NHANES offered an ideal dataset to focus on.

**Research Hypotheses**

Given the study objective, the following research question and hypotheses were developed to guide the project.

*Research Question*: Do results from the 2017-2018 NHANES support the indication that there is a link between mental health status and the presence of certain biochemical/nutritional markers?

*Null Hypothesis (H0)*: There is no evidence from the NHANES to suggest that a significant difference exists between certain biochemical/nutritional markers and mental health status.

*Alternate/Research Hypothesis (HA)*: There is evidence from the NHANES to suggest mental health status, in part, depends on the presence of certain biochemical/nutritional markers.

**Literature Review**

As noted in the introduction and author’s reasons for selecting the NHANES dataset, diet is a well understood component of overall health. For instance, cardiovascular disease has been linked with diets where consumption is high in saturated fats, refined sugars, and low in fruits and vegetables (Yu et. al., 2016, p. 1616). However as also noted in the introduction, relatively limited research in the United States has been conducted on whether diet plays a role in mental health directly. The author suspects that this is due to the historically perceived qualitative nature of emotional health conflicting with the empirical foundations of modern medical sciences.

Luckily, despite this conflict, there has been a movement towards viewing issues with emotional health as genuine medical conditions with which diet can have an impact on. In the studies that have been performed on the relationship between mental health and diet, results have indicated a significant link between the two. In one such study, promoting healthy dietary habits reduced anxiety in older adults (F. Masana, 2019). Furthermore, a study performed by Li et al (2017) writes, “the results of this meta-analysis suggest that healthy [dietary] pattern may decrease the risk of depression, whereas western-style [dietary patterns] may increase the risk of depression” (p.373). And retrospective studies like that performed by Khanna et al (2019) write, “Evidence is steadily growing for the relation between nutrition deficiencies, diet quality and mental health and for the efficacy and use of nutritional supplements to address deficiencies or as augmentation therapies. We advocate recognition of diet and nutrition as crucial factors in prevention and management of mental disorders” (para. 1).

The results and recommendations of these example studies have important implications for not only the treatment of mental disorders but for prevention as well. Therefore, this paper seeks to either corroborate or deny these findings through an independent study using the data outlined in the above and below sections.

**Research Design**

The results of the 2017-2018 NHANES is not a singular dataset; rather a number of datasets collected on the same individuals. Therefore, for the purposes of this project, only a subset of the available datasets has been selected for analysis because other available datasets were either not relevant to the study objective or inclusion would have resulted in a project beyond the scope of the assignment objectives. Ultimately, the following selected NHANES datasets were combined into a singular dataset: Dietary Interview – Total Nutrient Intakes First Day (DR1TOT\_J), Standard Biochemistry Profile (BIOPRO\_J), and Mental Health – Depression Screener (DPQ\_J).

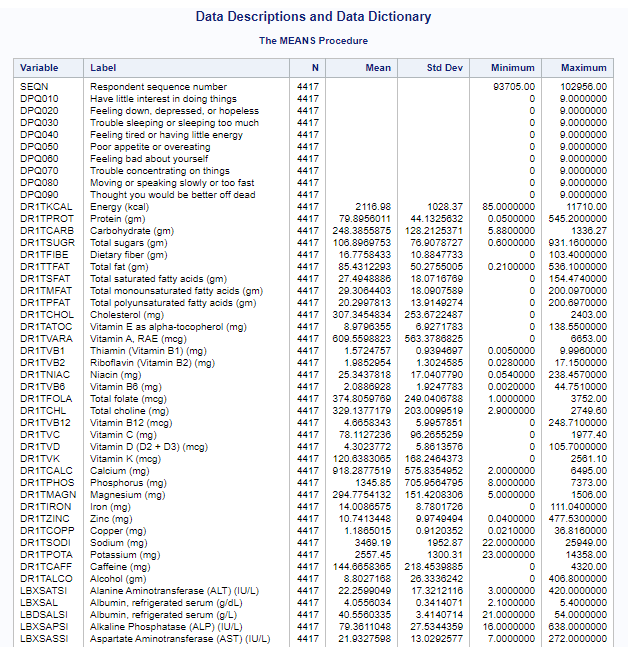
The DR1TOT\_J dataset is the results of an interview used to quantify the total nutritional intake of the participants 24 hours prior to the interview. 32 unique variables from this dataset were taken for the combined dataset. The BIOPRO\_J dataset is the results of blood tests taken from the same participants. Variables from BIOPRO\_J assist in the diagnosis of a variety of diseases related to organ function and metabolic performance. Details about how these variables are used to assess disease can be found [here](https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/BIOPRO_J.htm). 37 unique variables were taken from this dataset and entered into the combined dataset. Lastly, the DPQ\_J dataset is a depression screening questionnaire. This data is used to assess the frequency of depression symptoms during a 2-week period. 9 unique variables were taken from this dataset and entered into the combined dataset.

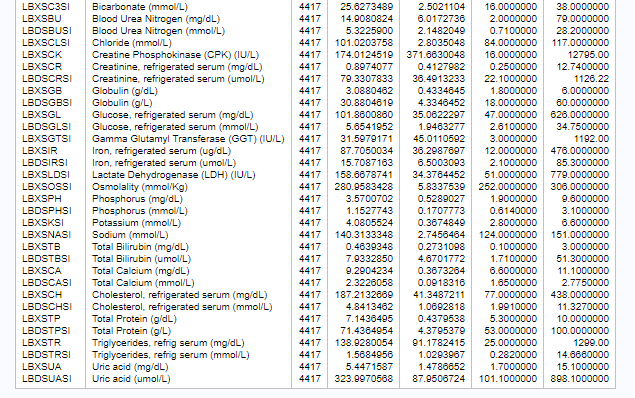
Figure 1 provides brief variable descriptions, along with basic descriptive statistics for the combined dataset. It should be noted that for all variables that start with “DPQ”, values are assigned positive integer numeric values that represented a scale of severity or indication of variable status; SEQN is a unique whole number value identifier used to identify each participant and also serves as the foreign key for combining the datasets; and all other variables are continuous numeric values with units of measurement defined under the Label column. For more detailed information about the DPQ\_J dataset variables, follow this [link](https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DPQ_J.htm#DPQ100). More detailed information for the DR1TOT\_J dataset variables can be found [here](https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DR1TOT_J.htm#WTDRD1). Similarly, additional details about the BIOPRO\_J dataset variables are located [here](https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/BIOPRO_J.htm).

Details about the methodology, analysis methods, and analysis tools as they relate to the research design are found in subsequent sections.

**Figure 1**

*Combined Dataset Dictionary with Data Names, Descriptions, and Summary Statistics*





*Note.* All variables that start with DPQ are from the DPQ\_J dataset. Similarly, all variables that start with DR1 are from the DR1TOT\_J dataset; and all variables that start with LB are from the BIOPRO\_J dataset.

**Methodology**

Few health studies in the United States include comprehensive nutritional intake, biochemical lab results, and depression screeners as data sources for the same participants. This means that the combined dataset offers a unique opportunity for a mixed methodology study where the interactions of quantitative biochemical/metabolic data can be viewed alongside with and related to the more qualitative emotional state/psychological data. This mixed method is beneficial because as O’Leary (2021) writes, “…mixed [methodology] approaches have the potential to be expansive… this allows mixed-methods research to: build a broader view by adding depth and insights to ‘numbers’ through inclusion of dialogue, narratives and pictures… [and] allow for triangulation” (p. 173).

**Methods & Tools**

The dependent variable for the testing the project hypotheses was DPQ020. DPQ020 is one of the DPQ\_J Mental Health – Depression Screener dataset variables. In summation, this variable asked participants the question, “Over the last 2 weeks, how often have you been bothered by the following problems: feeling down, depressed, or hopeless?”. Available responses were: 0 – Not at all, 1 – several days, 2- more than half the days, 3 – nearly every day, 7 – refused, 9 – Don’t know, and “.” – Missing. It should be noted that values of 7, 9, or “.” were removed from the dataset as they are essentially missing values. DPQ020 was selected as the dependent variable because it asked a direct mental health status question with clear and definitive answers. This variable was also already numerically coded making it easy to input into quantitative statistical tests despite being qualitative in nature.

A number of statistical tests, tools, and models proved useful in assessing the dependency between DPQ020 and the nutritional/biochemical variables. To begin, the majority of analysis took place in SAS Studio University Edition. This is because SAS Studio offers unparalleled simplicity and efficiency for performing comprehensive statistical and analytical tests. For instance, one of the tests performed on the dataset utilized the SAS PROC UNIVARIATE procedure code; which, produces a “…variety of statistics and graphs and is better suited to helping you discover important information about the distribution of each variable, such as whether: the data are approximately normally distributed or [if] there are outliers in the data (if so, where?)” (Elliott & Woodward, 2016, p. 224). Normality is an important test as it determines whether parametric or nonparametric tests may be selected.

After simple descriptive data tests, linear correlation tests were the next performed. These tests determined whether a linear relationship existed between the dependent variable and the independent variables. Correlation tests were selected because they offer the potential for important insights into the dependencies between variables; and they allow for the development of linear predictive models which could be used to determine mental health status when presented certain predictor markers.

Lastly, an ANOVA or Analysis of Variance statistical test was performed. This test compares the independent variable means between the different DPQ020 groupings. The ANOVA is an important test as it further indicates if there is a significant difference between DPQ020 groupings.

Lastly, the visualization tool Tableau was utilized to develop a basic data visualization that assisted with presenting the analysis results. Tableau was selected because it offers a user friendly platform for developing complex data visualizations in both a visually appealing and efficient way.

**SAS Data Import and Analysis Code**

libname sasfile '/folders/myshortcuts/SASDATA';

libname xptfile xport '/folders/myshortcuts/SASDATA/(x)Dietary Interview Total Nutrient Intakes 1st Day.XPT';

proc copy inlib=xptfile outlib=sasfile;

run;

libname sasfile '/folders/myshortcuts/SASDATA';

libname xptfile xport '/folders/myshortcuts/SASDATA/(x)Mental Health - Depression Screener.XPT';

proc copy inlib=xptfile outlib=sasfile;

run;

libname sasfile '/folders/myshortcuts/SASDATA';

libname xptfile xport '/folders/myshortcuts/SASDATA/(x)Standard Biochemistry Profile.XPT';

proc copy inlib=xptfile outlib=sasfile;

run;

DATA MIS581DATA;

MERGE MYSASLIB.DPQ\_J (KEEP= SEQN DPQ010 DPQ020 DPQ030

DPQ040 DPQ050 DPQ060 DPQ070 DPQ080 DPQ090)

MYSASLIB.DR1TOT\_J (KEEP= SEQN DR1TKCAL DR1TPROT

DR1TCARB DR1TSUGR DR1TFIBE DR1TTFAT DR1TSFAT

DR1TMFAT DR1TPFAT DR1TCHOL DR1TATOC DR1TVARA

DR1TVB1 DR1TVB2 DR1TNIAC DR1TVB6 DR1TFOLA

DR1TCHL DR1TVB12 DR1TVC DR1TVD DR1TVK DR1TCALC

DR1TPHOS DR1TMAGN DR1TIRON DR1TZINC DR1TCOPP

DR1TSODI DR1TPOTA DR1TCAFF DR1TALCO)

MYSASLIB.BIOPRO\_J (DROP= LBDSATLC LBDSGTLC LBDSTBLC);

BY SEQN;

RUN;

options missing='';

DATA MIS581DATA1;

SET MIS581DATA;

IF missing(DPQ010) then delete;

run;

DATA MIS581DATA2;

SET MIS581DATA1;

IF missing(DPQ020) or missing(DPQ040) or missing(DPQ060) or

missing(DPQ090) or missing(DR1TKCAL) or missing(LBXSASSI)

or missing(LBXSLDSI) or missing(LBDSTBSI) then delete;

run;

PROC MEANS DATA=MIS581DATA2;

Title "Data Descriptions and Data Dictionary";

RUN;

DATA Analysisdata;

SET MIS581DATA2;

IF DPQ020='.' or DPQ020=7 or DPQ020=9 then delete;

KEEP DPQ020 DR1TKCAL DR1TPROT DR1TCARB DR1TSUGR DR1TFIBE DR1TTFAT DR1TSFAT DR1TMFAT DR1TPFAT DR1TCHOL DR1TATOC DR1TVARA DR1TVB1 DR1TVB2 DR1TNIAC DR1TVB6 DR1TFOLA DR1TCHL DR1TVB12 DR1TVC DR1TVD DR1TVK DR1TCALC DR1TPHOS DR1TMAGN DR1TIRON DR1TZINC DR1TCOPP DR1TSODI DR1TPOTA DR1TCAFF DR1TALCO LBXSATSI LBXSAL LBXSAPSI LBXSASSI LBXSC3SI LBXSBU LBXSCLSI LBXSCK LBXSCR LBXSGB LBXSGL LBXSGTSI LBXSIR LBXSLDSI LBXSOSSI LBXSPH LBXSKSI LBXSNASI LBXSTB LBDSTBSI LBXSCA LBXSCH LBXSTP LBXSTR LBXSUA;

RUN;

PROC MEANS DATA=Analysisdata;

RUN;

PROC UNIVARIATE NORMAL PLOT DATA=Analysisdata;

RUN;

PROC CORR Data=Analysisdata SPEARMAN;

VAR DR1TKCAL--LBXSUA;

WITH DPQ020;

RUN;

PROC REG;

MODEL DPQ020=DR1TFIBE DR1TFOLA DR1TCHL DR1TVK DR1TMAGN DR1TCOPP DR1TPOTA LBXSAL LBXSCK;

TITLE "Multi-Regression Model Using Only Independent Variables that Showed Significant Linear Correlation";

RUN;

QUIT;

PROC REG;

MODEL DPQ020=DR1TKCAL--LBXSUA;

TITLE "Multi-Regression Model Using all Independent Variables";

RUN;

QUIT;

PROC ANOVA DATA= Analysisdata;

CLASS DPQ020;

MODEL DR1TFIBE DR1TFOLA DR1TCHL DR1TVK DR1TMAGN DR1TCOPP DR1TPOTA LBXSAL

LBXSCK = DPQ020;

MEANS DPQ020/TUKEY;

TITLE "ANOVA Severity of Depression for all Correlatively Significant

Variables";

RUN;

PROC ANOVA DATA= Analysisdata;

CLASS DPQ020;

MODEL DR1TFOLA=DPQ020;

MEANS DPQ020/TUKEY;

TITLE "ANOVA of Severity of Depression by Folate Intake";

RUN;

PROC ANOVA DATA= Analysisdata;

CLASS DPQ020;

MODEL LBXSAL=DPQ020;

MEANS DPQ020/TUKEY;

TITLE "ANOVA of Severity of Depression by Albumin Measurement";

RUN;

PROC ANOVA DATA= Analysisdata;

CLASS DPQ020;

MODEL DR1TFIBE=DPQ020;

MEANS DPQ020/TUKEY;

TITLE "ANOVA of Severity of Depression by Dietary Fiber Intake";

RUN;

PROC ANOVA DATA= Analysisdata;

CLASS DPQ020;

MODEL DR1TMAGN=DPQ020;

MEANS DPQ020/TUKEY;

TITLE "ANOVA of Severity of Depression by Magnesium Intake";

RUN;

PROC ANOVA DATA= Analysisdata;

CLASS DPQ020;

MODEL DR1TPOTA=DPQ020;

MEANS DPQ020/TUKEY;

TITLE "ANOVA of Severity of Depression by Potassium Intake";

RUN;

**Limitations**

Before presenting the analysis results, it should be noted that there were a number of limitations discovered within this study. First of all, the study population may not be representative of demographic distributions for the United States as a whole. For instance, the 2017-2018 NHANES documentation notes over sampling of certain ethnic populations; specifically, Hispanic individuals, Non-Hispanic black individuals, Non-Hispanic Asian individuals, Non-Hispanic white and other individuals below a poverty level, and Non-Hispanic white and other individuals that were 80 years or older (Centers for Disease Control and Prevention, 2021, para. 5).

In addition to this oversampling, certain traits such as age and sex are not accounted for in the analysis. Therefore, these missing variables could have an impact on the validity of the results presented in this study. However, the sole goal of the study was to either support or deny indications of the impact diet has on mental health; not to serve as a definitive conclusion for the relationship. Therefore, future studies would benefit by breaking out the participants into certain groups that control for these possibly confounding variables.

Furthermore, to account for differences in types of food intake, variables on the nutrition side look solely at macronutrient and micronutrient quantities rather than individual food items. This is because things like Vitamin C from an apple versus an orange are biochemically equivalent. However, looking to see if emotional health is related to certain food items (ie cheeseburgers versus spinach) would make for an interesting future study.

Lastly, there are notable limitations in the data variables. Particularly for variables related to the nutritional intake screener. In short, the nutritional intake values are limited to a single day of intake. Ultimately, this means that the values reported do not provide a very comprehensive view of the participant’s typical daily intake; and therefore severely limits the ability to identify nutrient patterns of intake that could influence the study objective.

**Ethical considerations**

In addition to noting study limitations, ethical considerations should also be addressed. Addressing privacy, security, and ethical concerns in data and health related studies really comes down to ensuring the individual’s identity is kept anonymous, participants are not harmed, data is secure from mal-intent, and that data participants are adequately informed with transparent data use policies. To this end, a number of actions have been taken to ensure the dataset used in this study meet these criteria. First, individuals are only identifiable by their unique numeric identifier. Any other identifying characteristics such as, location, name, or phone numbers, have been removed. Second, all data collected followed routine medical practices where applicable to ensure no participant was harmed. Third, participants are protected by an assurance of confidentiality which could result in a $250,000 fine if breached (National Center for Health Statistics, 2018, p. 2). Fourth, participants were informed in how their data could be used to study nutrition and health in the United States prior to participation. Lastly, analysis was performed with the intent of informational purposes only and is not intended to treat or diagnose any disease.

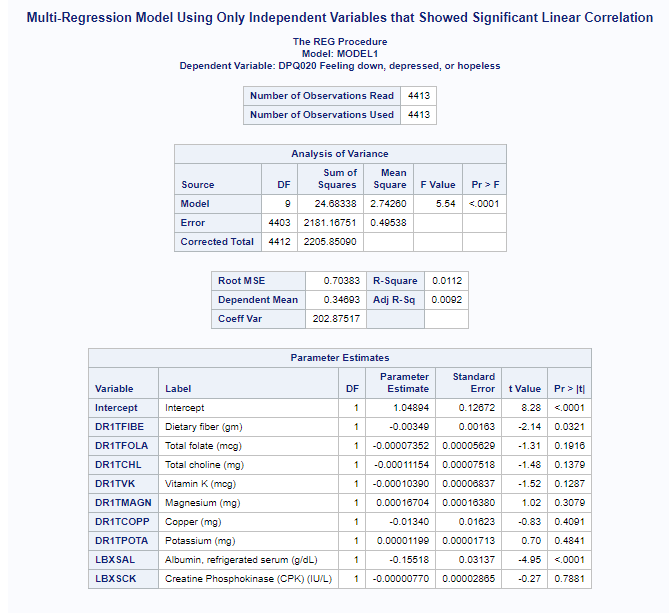
**Analysis Findings**

As mentioned in the Methods & Tools section, analysis began with assessing the variables for Normal Distribution utilizing the PROC UNIVARIATE code in SAS. Upon code execution, it was found that none of the variables were normally distributed because all “Tests for Normality” for each variable, the Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling tests, all showed p-values of less than 0.05. This guided the following linear correlation analysis to utilize the popular Spearman’s p-test; which is a non-parametric test and best utilized when the assumption for normalcy is not valid (Elliott & Woodward, 2016, p.284).

Results of the Spearman’s p-test all showed that relatively weak linear correlations between DPQ020 and the other independent variables existed. However, despite the weak correlations reported, the test did find that some of the correlations were significant with p-values of less than 0.0001. In particular, the variables for Dietary Fiber intake, Folate intake, Choline intake, Vitamin K intake, Magnesium intake, Copper intake, Potassium intake, measured Albumin, and measured Creatine Phosphokinase all showed a significant correlation. It should also be noted, that all significant correlations showed a negative linear relationship.

In response to the presence of significant correlative relationships, a multi-linear regression analysis was performed using DPQ020 as the dependent variable and the variables found to have a significant correlative relationship as the independent variables. Results of the multi-linear regression model can be found in Figure 2.

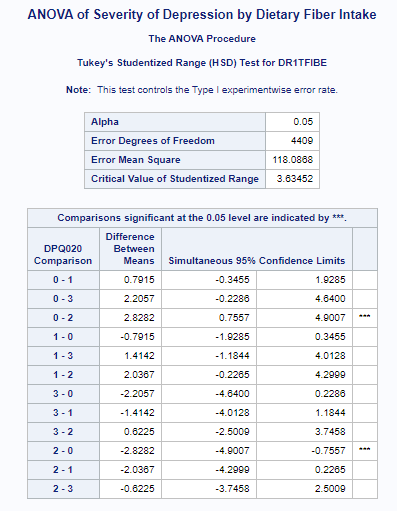
**Figure 2**

****

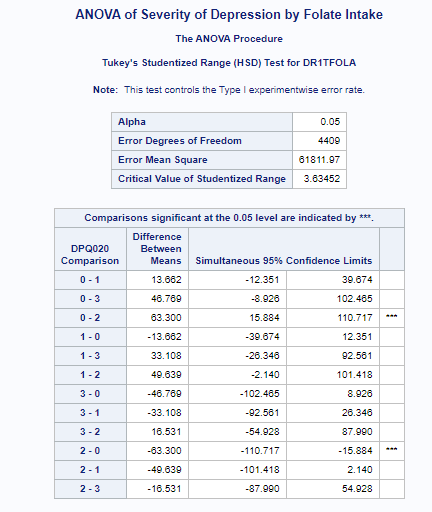
Results of Figure 2 show that both Dietary Fiber (DR1TFIBE) and Albumin (LBXSAL) had significant predictive relationships with the dependent variable, DPQ020. Furthermore, the p-value of less than 0.05 in the Analysis of Variance table also indicates the model is predictive of DPQ020. However, the R-Squared value is very low which indicates that the model does not predict DPQ020 very well.

As also mentioned in the Methods & Tools section, the correlative and regression model tests were then followed by several ANOVA tests. The ANOVA was performed despite the inability to assume normalcy because the sample size was sufficiently large to ignore departures from normalcy and because the critical assumption of sample independence was valid (Elliott & Woodward, 2016, p. 312). These ANOVAs compared the means of the DPQ020 variable groupings against all variables where significant correlative relationships were noted. Results of the ANOVAs that had significant findings of the ANOVAs can been found in the following Figures.

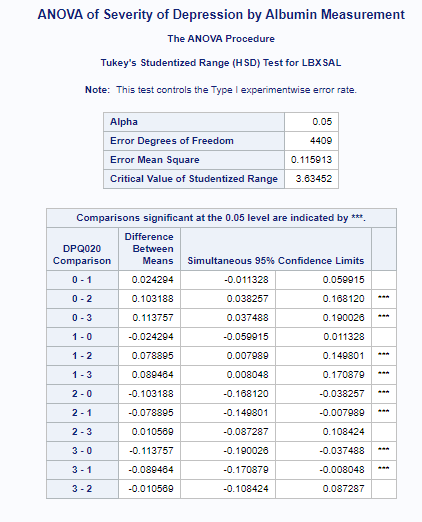
**Figure 3**

****

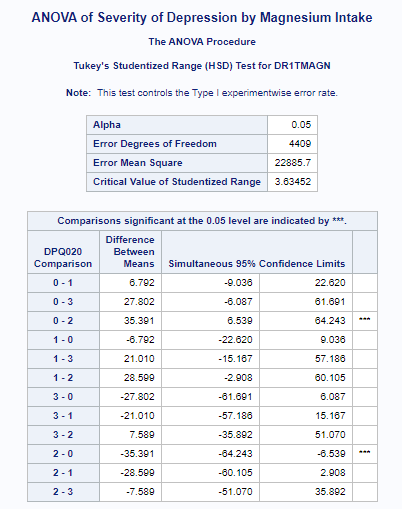
**Figure 4**

****

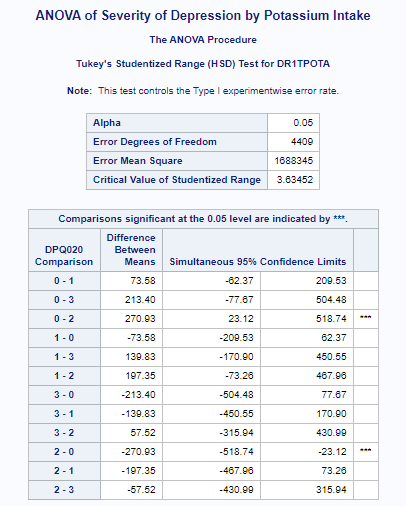
**Figure 5**

****

**Figure 6**

****

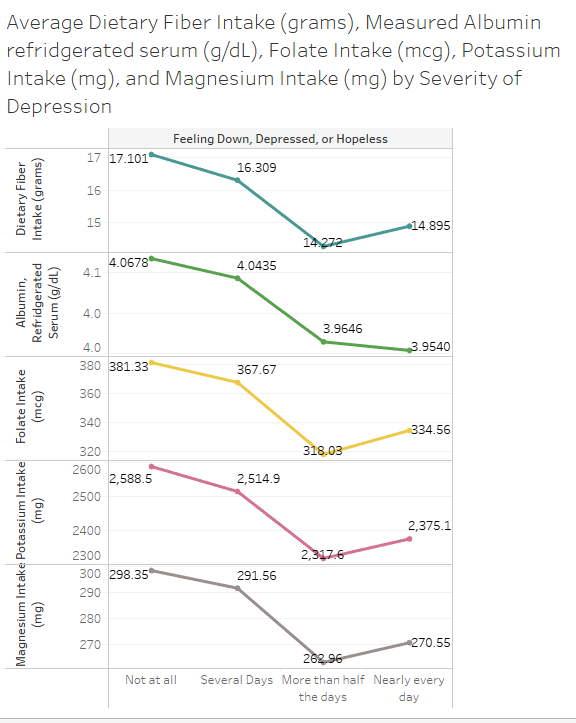
**Figure 7**

****

Figures 3, 4, 6, and 7 show that a significant difference exists between the mean fiber intake, folate intake, potassium intake, and magnesium for individuals that experienced no depression symptoms (0) and those that experienced symptoms more than half the days (2) prior to the screening. Comparatively, Figure 5 shows that a significant difference also exists between the mean measured Albumin values for individuals that experienced no depressions symptoms (0) and those that experienced symptoms more than half the days (2) or every day (3) prior to the screening. Figure 5 also notes that a significant difference in the measured Albumin means occurred between individuals that only experienced depression symptoms for several days (1) versus those that experienced symptoms more than half the days (2) and nearly every day (3).

Lastly, Tableau was utilized to help visualize any trends that were present for the significant correlative variables. These visualizations can be found in Figure 8.

**Figure 8**

****

**Conclusions**

Despite the poor multi-regression model performance, all tests ended up yielding results that found significant relationships between the depression values reported in DPQ020 and select biochemical/nutritional markers. In particular, the ANOVA from Figure 5 indicated that decreasing measured Albumin may be particularly indicative of mental health status. In light of these significant results, this study concludes that the null hypothesis should be rejected; therefore indicating that the results of the 2017-2018 NHANES may support a link between mental health status and certain biochemical/nutritional markers. However, due to the issues noted during the limitations section, this study’s conclusions should be met with a degree of skepticism.

**Recommendations**

SAS Studio and Tableau performed effectively for managing and analyzing the data presented in this study. These tools have significant capabilities that could be used to effectively support future studies. However, in the future, a number of controls should be implemented when studying the objective outlined in this project. Two main controls should include the following:

* Age and Sex should be accounted for
* A food log that includes all nutritional uptake for the same 2 weeks accounted for in the Depression Screener Questionnaire. This would help substantiate any relationships between mental health status and nutritional intake

In addition to the controls, different correlative and predictive models should be explored. For instance, a classification trees might yield better predictive results.

**Link to GitHub**

* User Profile: <https://github.com/charlie-seglem>
* Project Link: <https://github.com/charlie-seglem/MIS581-Portfolio-Project-Does-Nutrition-Influence-Incidences-of-Depression>

**References**

Auestad, N., & Fulgoni, V. L., 3rd. (2015). What current literature tells us about sustainable diets: emerging research linking dietary patterns, environmental sustainability, and economics. *Advances in Nutrition (Bethesda, Md.)*, *6*(1), 19–36. <https://doi.org/10.3945/an.114.005694>

Centers for Disease Control and Prevention. (n.d.). *About NCHS - The NCHS Mission*. Centers for Disease Control and Prevention. <https://www.cdc.gov/nchs/about/mission.htm>.

Centers for Disease Control and Prevention. (2020, March 30). *About NCHS - Celebrating 50 Years*. Centers for Disease Control and Prevention. <https://www.cdc.gov/nchs/about/50th_anniversary.htm>.

Centers for Disease Control and Prevention. (2020, October 6). *About NCHS - The NCHS Budget*. Centers for Disease Control and Prevention. <https://www.cdc.gov/nchs/about/budget.htm>.

Centers for Disease Control and Prevention. (2021, April 9). *NHANES 2017-2018 Overview*. Centers for Disease Control and Prevention. <https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overview.aspx?BeginYear=2017>.

Elliott, A. C., & Woodward, W. A. (2016). *Sas essentials: mastering Sas for data analytics*. Wiley.

F. Masana, M., Tyrovolas, S., Kollia, N., Chrysohoou, C., Skoumas, J., Haro, J. M., … B. Panagiotakos, D. (2019). Dietary Patterns and Their Association with Anxiety Symptoms among Older Adults: The ATTICA Study. *MDPI: Nutrients*, *11*(6), 1250. <https://doi.org/10.3390/nu11061250>

Fulgoni, V. L., & Buckley, R. B. (2015). The Contribution of Fortified Ready-to-Eat Cereal to Vitamin and Mineral Intake in the U.S. Population, NHANES 2007-2010. *Nutrients*, *7*(6), 3949–3958. <https://doi.org/10.3390/nu7063949>

Khanna, P., Chattu, V. K., & Aeri, B. T. (2019). Nutritional Aspects of Depression in Adolescents - A Systematic Review. *International journal of preventive medicine*, *10*, 42. <https://doi.org/10.4103/ijpvm.IJPVM_400_18>

Kim, K., Vance, T. M., Chen, M.-H., & Chun, O. K. (2018). Dietary total antioxidant capacity is inversely associated with all-cause and cardiovascular disease death of US adults. *European Journal of Nutrition*, *57*(7), 2469–2476. <https://doi.org/10.1007/s00394-017-1519-7>

Li, J., Lee, D. H., Hu, J., Tabung, F. K., Li, Y., Bhupathiraju, S. N., Rimm, E. B., Rexrode, K. M., Manson, J. E., Willett, W. C., Giovannucci, E. L., & Hu, F. B. (2020). Dietary Inflammatory Potential and Risk of Cardiovascular Disease Among Men and Women in the U.S. *Journal of the American College of Cardiology*, *76*(19), 2181–2193. <https://doi.org/10.1016/j.jacc.2020.09.535>

Li, Y., Lv, M.-R., Wei, Y.-J., Sun, L., Zhang, J.-X., Zhang, H.-G., & Li, B. (2017). Dietary patterns and depression risk: A meta-analysis. *Psychiatry Research*, *253*, 373–382. <https://doi.org/10.1016/j.psychres.2017.04.020>

Murray, C. J. L. (2018). The State of US Health, 1990-2016 Burden of Diseases, Injuries, and Risk Factors Among US States. *Journal of American Medical Association*, *319*(14), 1444–1472. <https://doi.org/doi:10.1001/jama.2018.0158>

National Center for Health Statistics. (2018). *2017 NHANES Confidentiality Brochure*. <https://www.cdc.gov/nchs/nhanes/participant/2017_NHANES_English_Confidentiality_Brochure_ERB.pdf>.

O'Leary, Z. (2021). *The Essential Guide to Doing Your Research Project* (4th ed.). SAGE PUBLICATIONS.

Wikimedia Foundation. (n.d.). *National Center for Health Statistics*. Wikipedia. <https://en.wikipedia.org/wiki/National_Center_for_Health_Statistics>.

Yu, E., Rimm, E., Lu Qi, Rexrode, K., Albert, C. M., Qi Sun, Willett, W. C., Hu, F. B., & Manson, J. E. (2016). Diet, Lifestyle, Biomarkers, Genetic Factors, and Risk of Cardiovascular Disease in the Nurses’ Health Studies. *American Journal of Public Health*, *106*(9), 1616–1623. <https://doi.org/10.2105/AJPH.2016.303316>