**Statistical Analysis Plan**

What are the Key Social Indicators for Depression based on Age Groups?

**Boston University**

**Master of Science in Statistical Practice (MSSP)**

Colin Clapham

**Project Background**

**Relevant and Situational Medical Context**

Depression is a mental health disorder that affects the daily lives of those suffering from the illness. In the US alone, it is estimated that over three million people are diagnosed with depression every year. While the disorder can cause significant impairment to daily life, it has become very common and treatments are available.

Causes of depression have been linked to biological, psychological, and social sources of distress. Trauma from one of these categories can lead to many symptoms, most commonly a depressed mood combined with a lack of interest in daily activities. Most often, treatment involves medication and/or therapy.

**Research Objectives**

**Why am I Performing this Research?**

Depression is a disease that affects all ages from early teens all the way to the end of life. The disorder also goes undiagnosed for many with many people refusing care or even receiving improper diagnosis. It is extremely important to understand the factors that lead to the breakdown of the psyche.

Through this research, I intend to dig deeper into the social sources of distress and their effect on patients across multiple age brackets. There may be social constructs which affect people with depression in different ways based on their age category. I intend to look at variables such as education levels, family status, activity levels, etc. and analyze their contribution to whether someone was diagnosed with depression.

**What is the Principal Outcome I want to Achieve?**

For the purposes of this research I will be fitting a multi-level model with each level corresponding to a pre-determined age bracket. The variables outlined below will be used to predict the likelihood of a depression diagnosis:



Survey Year – year in which the survey was conducted

Sex – indicator of male or female

Age – age at time of survey (will be broken into 10 year brackets for multi-level modeling)

CDCMSTAT – marriage status on a scale (Separated, Divorced, Married, Single, Widowed, Unknown)

Parents – status of parents on a scale (Mother only, Father only, Father and Mother, Neither, Unknown)

LAHCA17 – indicator of a current diagnosis of depression

EDUC1 – highest education level achieved on a scale indicating years of schooling

WRKHRS2 – continuous variable indicating hours worked per week

**Demographics to Control**

Based on public research, there is a fairly even split in age groups in terms of diagnoses of depression. It will be important to determine the breakdown of ages in the data I have pulled. Overall, I will need to put each variable in context (ie – younger people will not be married, older people may be retired and thus won’t be working).

According to the Anxiety and Depression Association of America, the estimated average onset age of Persistent Depressive Disorder (PDD) is around 31 years old. In addition, women are almost twice as likely to receive a diagnosis of depression compared to men. I will need to keep these factors in mind when looking at subsets of my data, specifically the patients with positive diagnoses of depression.

**Data Collection and Cleaning**

All survey data used in my analysis was collected from the CDC’s National Center for Health Statistics. Data from the National Health Institute Survey (NHIS) has collected and published on a yearly basis dating back to 1997. A major challenge in reading this data was the different formats used in the published data. The overall goal was to read everything and perform all analysis in R.

For the purposes of my analysis, I looked to read all survey data over a ten year period starting in 2007 and ending in 2016. While the survey data for 2015 and 2016 were published in an easily readable csv format, all other years were published in formats that were only able to be read using SAS. Using code published for the NHIS combined with code written to export files into csv format, I was able to easily read all survey data into R.

Initially, once all survey data was read into R I had just under 700,000 observations with over 600 variables for most years. I went through and selected the variables I wished to use in my analysis (ie – those pertaining to social factors and depression). In the end, I was left with over 76,000 observations and the 9 variables described above.

**Output**

**Format of the output I plan to provide:**

**Part I: Data Description**

Description of the data source and a summary of the data I will be using

**Part II: EDA**

1. Histogram of positive versus negative diagnosis for all survey respondents
2. Histogram of gender breakdown for patients diagnosed with Depression (compared to those without depression)
3. Histogram of age breakdown for patients diagnosed with Depression (compared to those without depression)
4. Histogram of marriage status for patients diagnosed with Depression (compared to those without depression)

**Part III: Build a multi-level model to predict the likelihood of a depression diagnosis based on age bracket**

1. Interpret results to reveal what social factors contribute more to a positive depression diagnosis based on age

**Part IV: Summary**

**Sources**

Depression Facts - <https://www.gstatic.com/healthricherkp/pdf/clinical_depression.pdf>

<https://adaa.org/about-adaa/press-room/facts-statistics>

Age Bracketing -<https://www.medicinenet.com/script/main/art.asp?articlekey=22674>

Variable Definitions - <ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2016/personsx_summary.pdf>

Variable Summary - <ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2016/personsx_layout.pdf>