Association between screen time and depression: a population-based study using the 2011-2012 national health and nutrition examination sample

CHL7001HY1 – Final Project

Name: Tian Han (Eric) **Guan**

Student ID: 998978058

Department: Statistics

Program: MSc

Email: tianhan.guan@mail.utoronto.ca

Abstract

Growing evidence starts to show depression and mental health issues have been rising among Americans and will become one of the leading diseases by the year 2030. In recent years, epidemiological researchers have found that personal lifestyle is an important risk factor of experiencing depression problems for Americans in various groups. The primary purpose of this study is to replicate the results from the research paper *Association between screen time and depression among US adults by K.C. Madhava et al.* which showed that there is an association between moderate or severe depression with more time spent on TV watching and computer use using the 2011-2012 NHANES sample data. In addition, this study also provides additional analyses and conclusions on the research question using various statistical methods. The survey package from R is used to perform all the statistical analyses and the ggplot2 package from R is used for all the data visualization analyses in this study.

Introduction

Depression (also known as major depressive disorder) is a common mental health illness that can negatively affect a person's thoughts and behaviours. According to the Centers for Disease Control (CDC), more than 9% of Americans suffer from depression related mental health symptoms such as hopelessness and despondency, which can further lead to various health risks such as diabetes and cardiovascular diseases. As a result, it is essential to understand the association between various risk factors and depression so that better management can be done efficiently.

The research question of this study is to investigate the association between screen time (the amounts of time that people spent on TV watching and computer use outside school or work) with depression, by considering other variables including gender, education level, age, race, BMI, and poverty class. This research question has been studied by *K.C. Madhava et al.* (*2017)* in their paper *Association between screen time and depression among US adults* which used the 2011-2012 National Health and Nutrition Examination survey (NHANES) sample data where it concluded that there was an association between people who spent more than 4 hours per day on TV/computer with experiencing moderate or severe depression. The goal of this study is to 1) replicate the results from the original research paper and 2) perform extra analyses to further support the conclusions and develop more insights on the research question.

This report is organized in the following structure. The first section discusses the sampling design used by the 2011-2012 NHANES including both the stages of the design as well as some insights on additional sampling techniques such as sample weight and oversampling. The second section derives all the study variables and explains the data building methodologies. Following that, the third section applies various data visualization techniques to draw some preliminary conclusions on the sample data. The fourth section performs both weighted and unweighted analyses for the same model and compares the estimates and their standard errors along with the results from the original research paper. The fifth section uses logistic regression to further analyze the association between depression and screen time using odds ratios and confidence intervals. The final section summarizes all the findings and suggests some limitations of the study as well as further study recommendations on this research topic.

1. Data collection of the study

The study of this report is conducted using the sample from the 2011-2012 National Health and Nutrition Examination survey (NHANES), a large cross-sectional U.S. survey conducted by the Centers for Disease Control (CDC) and National Center for Health Statistics (NCHS).

1.1 Study population

The NHANES contains the health examination data for a nationally representative sample of the non-institutionalized civilian resident of the U.S. population and excludes information about persons in custody or supervised care, active military personnel, families living abroad, and any other U.S. citizens residing outside of the 50 states and District of Columbia.

1.2 Sampling frame and Stratification

The frame of NHANES includes all counties in U.S. where 2,846 PSUs are formed from 3,100 counties and these PSUs are grouped into 13 major strata* based on grouping of states and urban-rural population distributions of the PSUs.

*The 2011-2012 NHANES sample is part of the 2011-2014 NHANES sample. For the 4-year sample, each major strata is further divided into 4 minor strata and 1 PSU is selected from each minor stratum for each year's sample (this approach allows the multiyear samples to be evenly distributed in terms of geographical regions and population characteristics). For this study, only the major strata are considered and will be used for variance estimation purposes.

1.3 Sampling design

The 2011-2012 NHANES uses a four-stage complex sampling design:

Stage 1: Selecting PSUs from a frame of all U.S. counties

- In most of the cases, PSUs are counties, while in a few circumstances, a PSU is the combination of some adjacent counties in order to make sure the PSU has a large enough

sample size. The 2011-2012 NHANES sample selects 2 distinct PSUs from each of the 13 major strata and the major strata are used as the strata for variance estimation. In total, 26 distinct PSUs were obtained in the sample.

- The PSUs were selected with probability proportional to size (PPS) within each stratum, in this case, size is the population in each county or combination of adjacent counties.

Stage 2: Selecting area segments from each PSU

- Each PSU was divided into multiple area segments (e.g. city blocks) which represent census blocks or combinations of blocks.
- The area segments were selected using the PPS method, in this case, size is the population in each area segment.
- The sample was designed so that each PSU and area segment has approximately equal sample size.
- On average, 1 PSU contains 24 area segments.

Stage 3: Selecting households from each area segment

- Within a given PSU, after the selection of area segments, a listing of all households was obtained and a subsample of the households was selected.
- The sampling rate of the households was set to produce an approximately equal probability sample of the households within each area segment.

Stage 4: Selecting persons within each household

- A list of all eligible individuals in a given household was obtained and a subsample of individuals was selected based on their characteristics (for example age, gender, race, origin, income, etc.).
- The sampling rate of the individuals was set to maximize the average number of sampled individuals per household.

The following table summarizes the sampling design used in the 2011-2012 NHANES:

Stage Number	What is being selected?	Sampling method	Number of units selected (in total)
1	PSUs (counties)	PPS	26
2	Area segments	PPS	720
3	Households	Sampling rate	23,000
4	Individuals	Sampling rate	10,000

Overall, use of the above multistage sampling design ensures that some fixed target sample sizes were achieved for each domain which was defined by individual characteristics such as gender, age, race, origin, and income status.

1.4 Sample weight and oversampling

In the NHANES sampling design, each individual has a sample weight assigned which represents the number of people in the study population in NHANES. The design uses unequal probability of selection (which means the sample weights are different for each sampled individual) in order to produce an unbiased estimate for the nation.

Oversampling is a common approach in large survey sampling designs which is used to adjust the distribution of subsamples of a sample dataset, so that the bias in estimations can be corrected. Here, NHANES uses oversampling to sample more individuals from certain subgroups for particular health interest so that the precision of estimates about the health information for these subgroups can be improved. For example, the following subgroups were being oversampled for the 2011-2012 NHANES (mostly based on race):

- Hispanic Americans
- Non-Hispanic black Americans
- Asians
- Low-income white Americans (at or below 130% of the poverty level)
- Adults of age 80 or over

2. Variables used in the study

2.1 Data exclusion

In this study, the following samples are excluded from the analysis:

- Individuals who were under the age of 20
- Individuals who did not respond to the mental health depression screen questionnaire (NAs to all questions)
- Individuals who had missing values (one or more NAs) in the demographic information dataset
- Individuals who had missing values (one or more NAs) for the physical activity questionnaire about TV watching/computer use
- Individuals who had missing values for their BMIs in the BMI dataset

The following datasets and variables have been used in the study (all datasets can be obtained from the Centers for Disease Control (CDC) website under the NHANES 2011-2012 section).

2.2 Demographics data - Demographic Variables & Sample Weights (DEMO_G)

The following variables are being used:

Age: individuals with age under 20 years are excluded from this study, otherwise, the age variable is categorized into 4 levels:

1) 20-35 years (coded 1)
2) 36-50 years (coded 2)
3) 51-65 years (coded 3)
4) >65 years (coded 4)

Gender:

1) Male (coded 1)
2) Female (coded 2)

Education is categorized into 2 levels based on the survey response below (note: in this study, score 7 and 9 are considered to be missing values (same as NA), although this is not being mentioned explicitly in the original research paper).

1) Less than high school/GED (coded 1)
2) High school graduate or more (coded 2)

Code or Value	Value Description from the survey
1	Less than 9th grade
2	9-11th grade (Includes 12th grade with no diploma)
3	High school graduate/GED or equivalent
4	Some college or AA degree
5	College graduate or above
7	Refused
9	Don't Know
NA	Missing

Race/ethnicity is categorized into 4 levels:

1) Non-Hispanic White (coded 1)
2) Non-Hispanic Black (coded 2)
3) Hispanic (coded 3)
4) Other race (coded 4)

Poverty (based on the income-to-poverty ratio) is categorized into 2 levels:

1) Yes - with an income-to-poverty ratio below 1 (coded 1)
2) No - with an income-to-poverty ratio above or equal to 1 (coded 2)

2.3 Examination data - Body Measures (BMX_G)

- Body Mass Index (BMI) = body weight/body height and is expressed in kg/m^2, the corresponding variable in the dataset is *BMXBMI*
- Individuals with missing BMI values are excluded from the study
- The **BMI** variable is grouped into 4 categories based on the following criteria:

BMI (kg/m^2)	BMI Group
<18.5	Underweight (coded 1)
18.5 to 24.9	Normal weight (coded 2)
25 to 29.9	Overweight (coded 3)
>30	Obese (coded 4)

2.4 Questionnaire data - Physical Activity (PAQ_G)

The physical activity questionnaire includes questions related to daily activities, leisure time activities, and sedentary activities. In this study, only the responses from the following 2 questions are used to determine the **screen time** variable:

I. Average number of hours spent on watching TV or videos per day over the past 30 days (PAQ710)
II. Average number of hours spent on using computer or playing computer games (outside of work or school) per day over the past 30 days (PAQ715)

The following table illustrates the meaning of each response from the survey:

Score of each question	Corresponding average time spent on TV or computer
0	Less than 1 hour
1	1 hour
2	2 hours
3	3 hours
4	4 hours
5	5 hours or more
8	{You do/SP does} not watch TV or videos
77	REFUSED

99	DON'T KNOW
NA	Missing

Note:

1. In this study, score 8, 77, and 99 are considered to be missing values (same as NA) (note: this is not being mentioned explicitly in the original research paper).
2. Individuals who did not have complete responses to the 2 questions (i.e. one or more NAs) are excluded from the study.
3. The *total screen score* is the sum of the scores of the 2 questions (which can range from 0 to 10 on an integer basis).

In this study, the variable **screen time** is a binary variable that is built using the following criteria:

Total screen score	**Screen time**
0-4	Low (coded 1)
5-10	High (coded 2)

Note: in the original research paper, the authors categorized the variable based on scores < 4 hours and scores > 4 hours, but did not mention the condition when scores = 4 hours (as this is a discrete variable).

2.5 Questionnaire data - Mental Health - Depression Screener (DPQ_G)

Depression was measured using the Patient Health Questionnaire (PHQ-9), a 9-question based self-report that asked questions about the frequency of symptoms of depression over the past 2 weeks. The questionnaire also has a final question to assess the overall impairment of the depressive symptoms but is not being used in this study.

The following table illustrates the meaning of each response from the survey:

Score of each question	Corresponding frequency of symptoms of depression
0	Not at all
1	Several days
2	More than half the days
3	Nearly every day
7	Refused
9	Don't know
NA	Missing

Note:

1. In this study, score 7 and 9 are considered to be missing values (same as NA). (Note: this is not being mentioned explicitly in the original research paper)
2. The *PHQ-9 score* is the total of the scores of all 9 questions (which can range from 0 to 27 on an integer basis).

In this study, the response variable **depression** is a binary variable that is built using the following criteria: (note: when adding up the scores, NAs are treated as 0s)

PHQ-9 score	Depression
0-9	No or mild (coded 1)
10-27	Moderate or severe (coded 2)

Note: in the original research paper, the authors categorized the variable based on scores < 9 and scores of 10 or more, but did not mention the condition when scores = 9 (since this is a discrete variable).

The following table summarizes all the variables used in this study:

In total, 2,969 samples are obtained in the final dataset.

Variable name	Type of variable	Code/value
Depression	Factor	1/2
Gender	Factor	1/2
Age	Factor	1/2/3/4

Education	Factor	1/2
Race	Factor	1/2/3/4
Poverty	Factor	1/2
BMI	Factor	1/2/3/4
Screen_time	Factor	1/2

Remarks about the obtained sample dataset:

In the original research paper, 3,021 samples were obtained after filtering and merging the datasets from NHANES (versus 2,969 samples obtained in this study). The difference may be caused by the following reasons:

- In this study, individuals who had missing values for their demographic information and BMIs are being removed. In the original research paper the authors only mentioned the removal of individuals who did not have complete demographic information, but did not mention anything about individuals who had missing BMI values.
- For various survey questions (e.g. frequency of depression, time spent on TV/computer, etc), there are some responses of either refused or do not know, in this study, these types of response are treated as missing (NA) values. In the original research paper, the authors did not mention how they treated such responses.

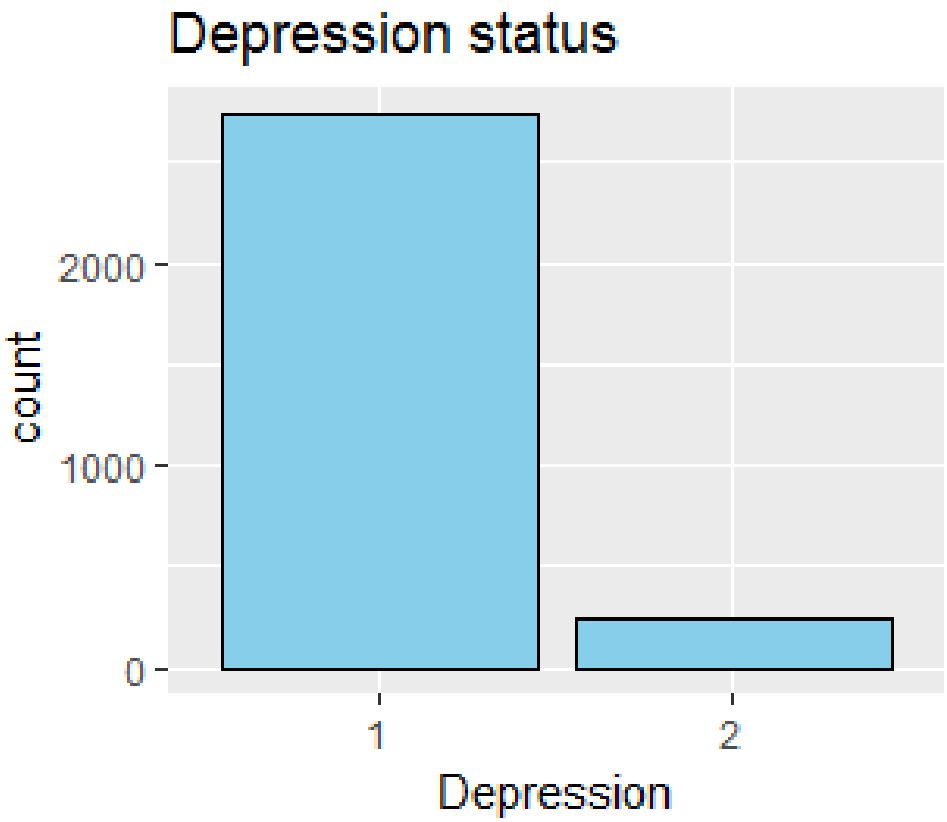
In addition to the above data filtering methodologies, other reasons such as programming errors and/or modification of the NHANES datasets may also contribute to the difference between the obtained sample size in this study and the sample size used in the original paper. However, since the difference is not significantly large and the data building methodologies are reasonable in this study, it is believed that the obtained sample dataset is a reasonable dataset to perform analysis and to be compared with the results of the original research paper.

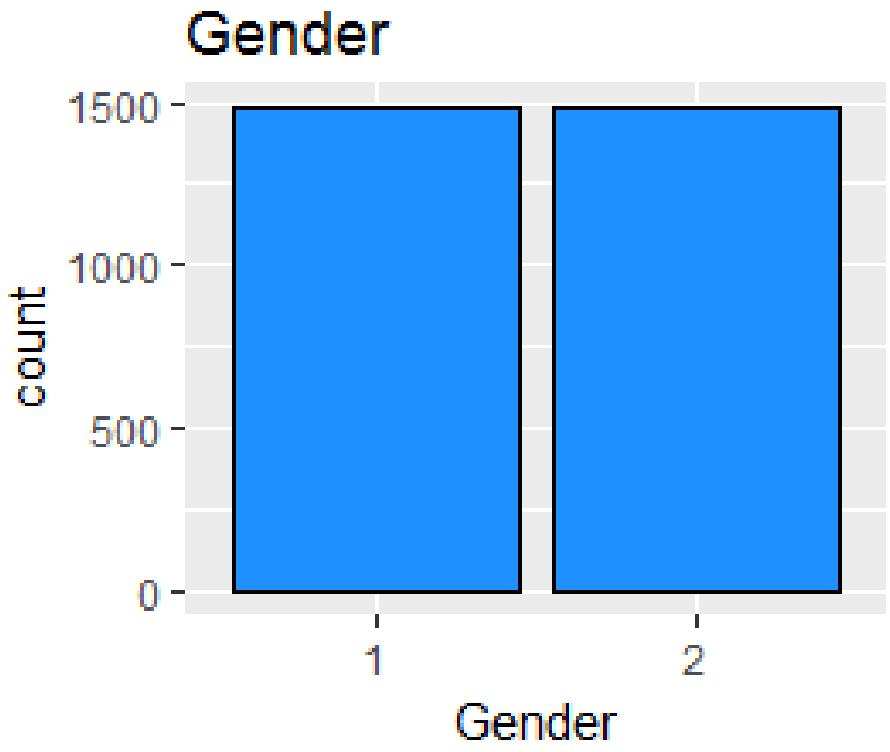
3. Exploratory Data Analysis of the sample dataset

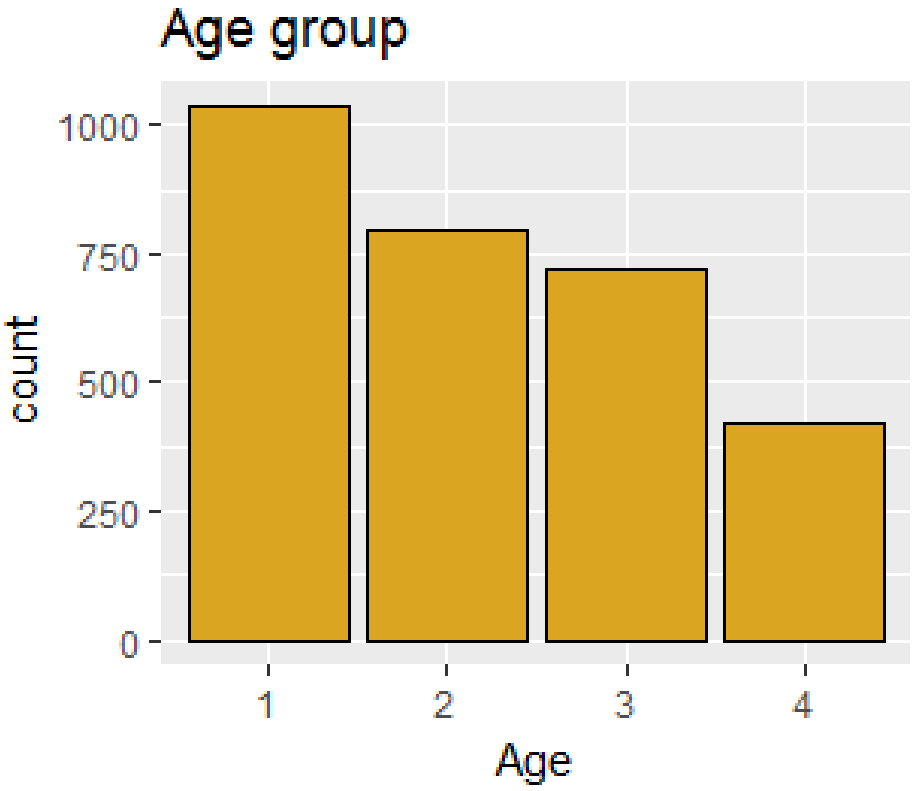
The next step after creating the final dataset is to have a brief look at each of the variables and the whole dataset using exploratory data analysis. The following section includes various graphical summaries of the dataset.

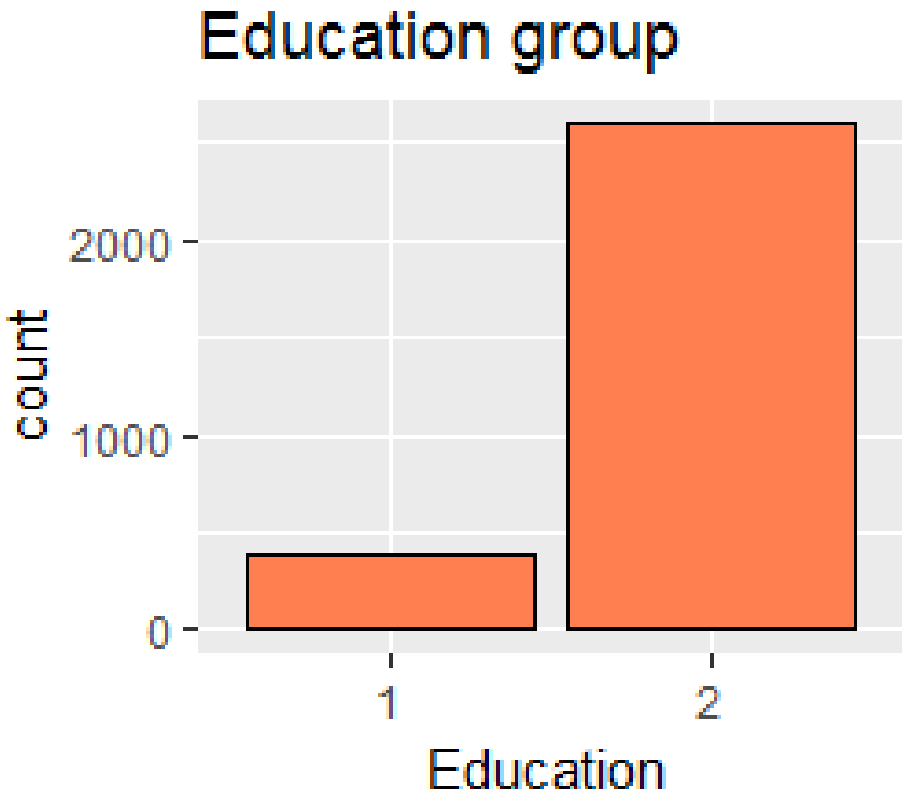
3.1 Distribution of factors within each variable

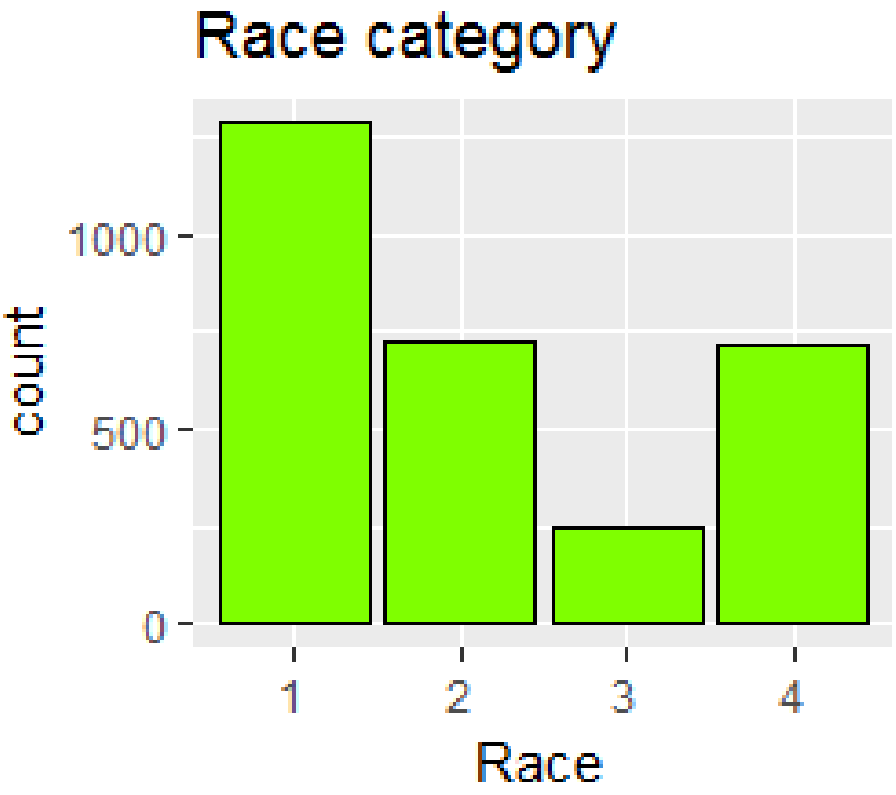
The following bar plots are used to explore the distribution of different factors/levels within each variable.

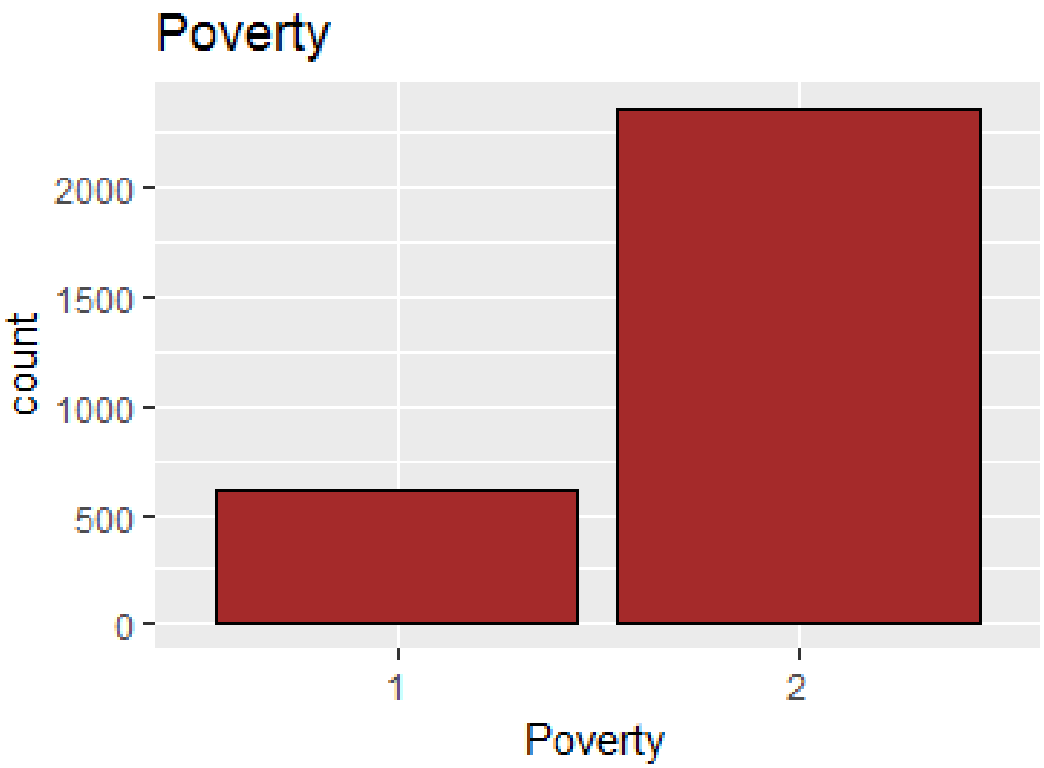


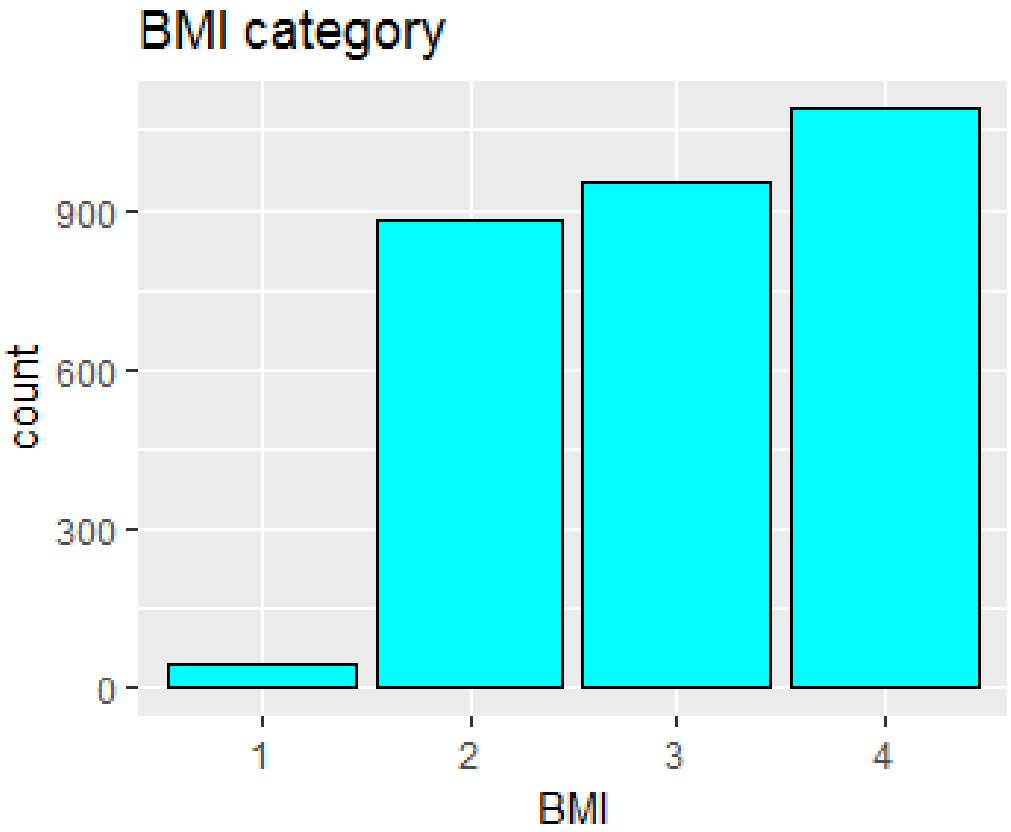

Gender
1500
1000
500
0
count
1
2
Gender

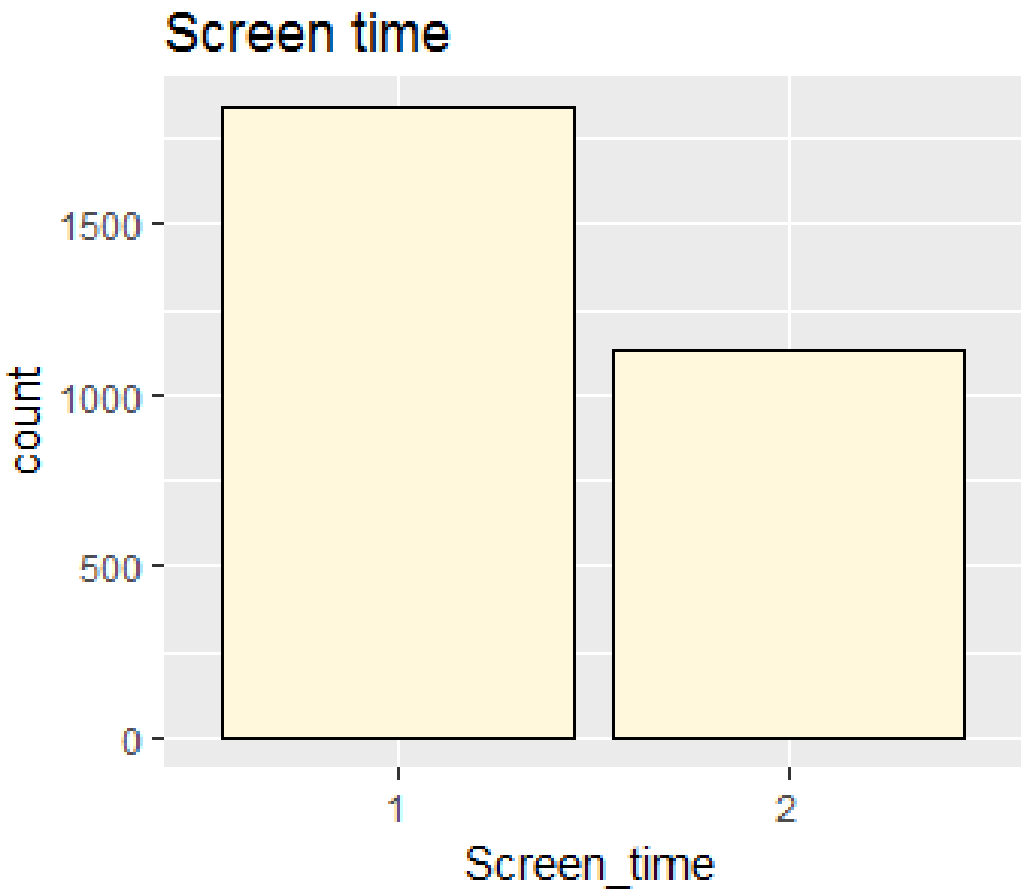

Age group
1000
750
500
250
0
count
1
2
3
4
Age


Education group
2000
1000
0
count
1
2
Education


Race category
1000
500
0
count
1
2
3
4
Race


Poverty
2000
1500
1000
500
0
count
1
2
Poverty


BMI category
900
600
300
0
count
1
2
3
4
BMI



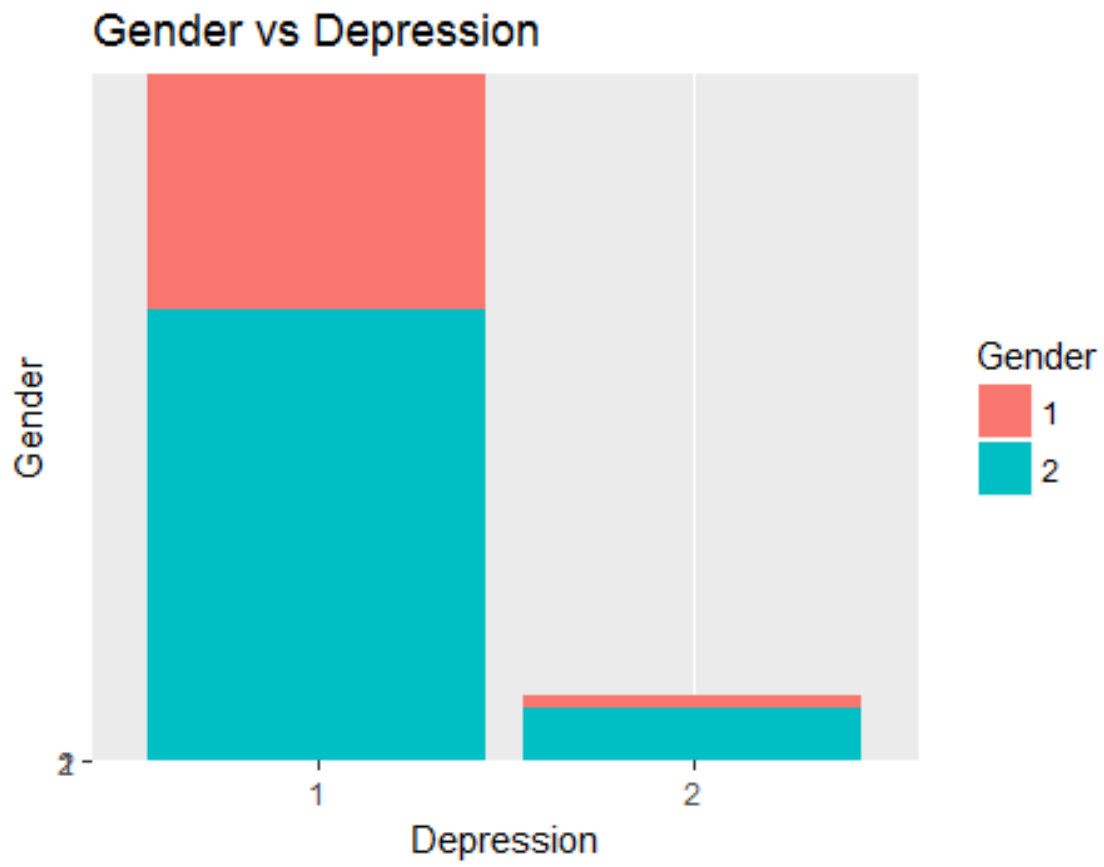
Comments on the above graphs

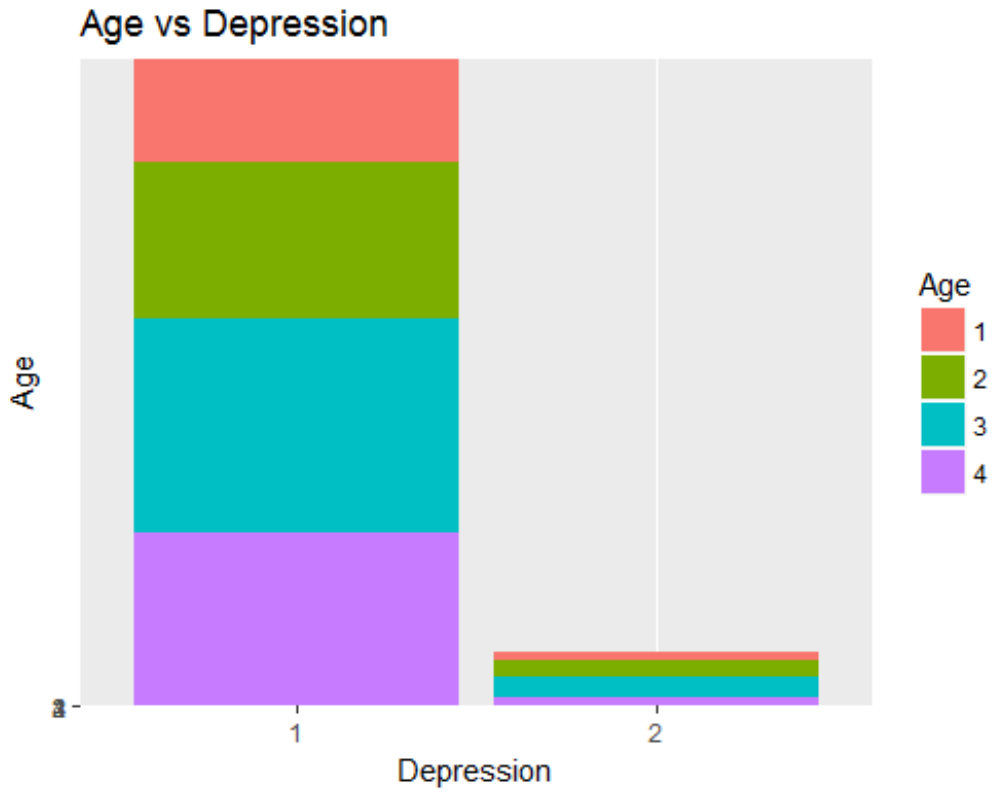
From the bar plots, the following observations can be made about the sample:

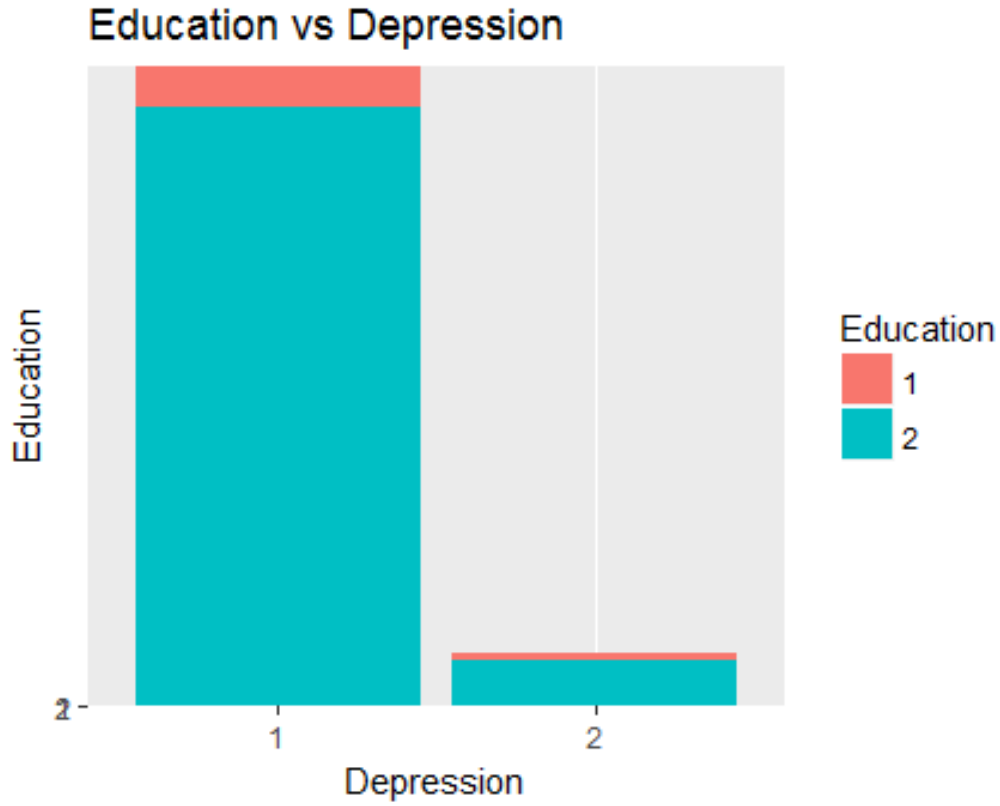
- Majority of the sampled individuals had no or mild depression
- The sample is well-balanced between both genders
- Majority of the sampled individuals were in their middle ages, and had at least high school level education
- Non-Hispanic white people were the most selected race in the sample
- Most of the sampled individuals were above the poverty threshold, however, there were also many individuals in the sample who were below the poverty threshold
- A large significant portion of the sampled individuals were in the 'unhealthy' category (with BMI overweight or obese)
- A large significant portion of the sampled individuals had high screen time on either TV or computer

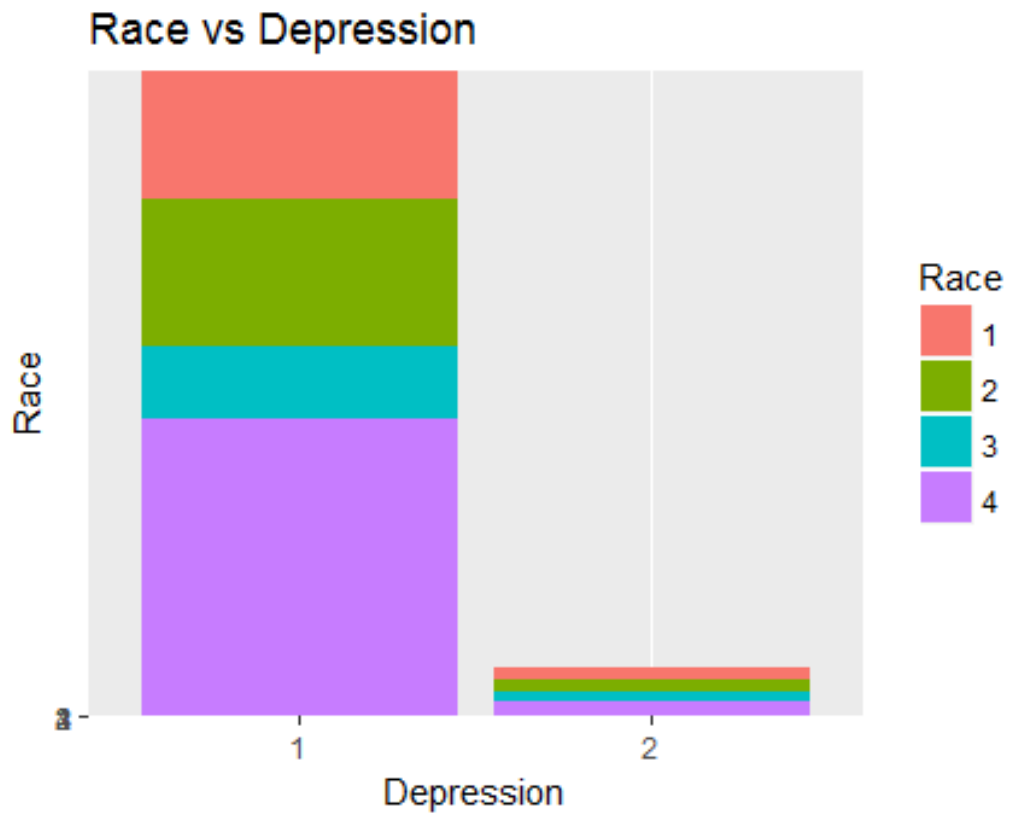
3.2 Association between depression and other study variables

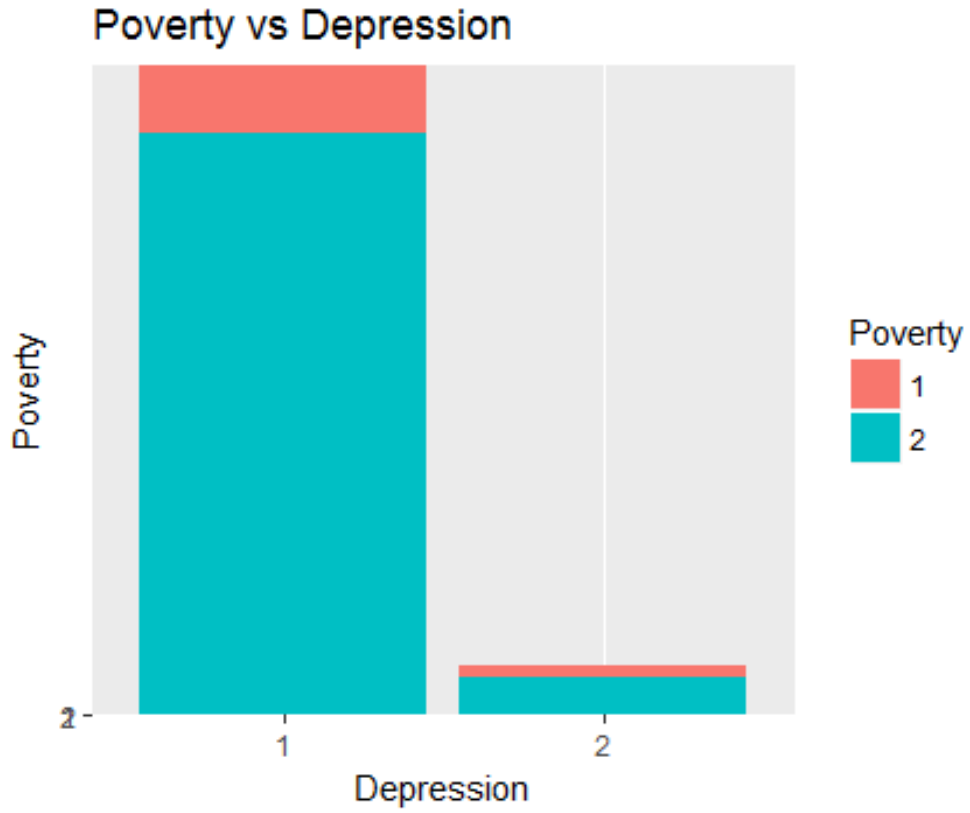
The following stacked bar plots are used to explore the association between depression and other variables:

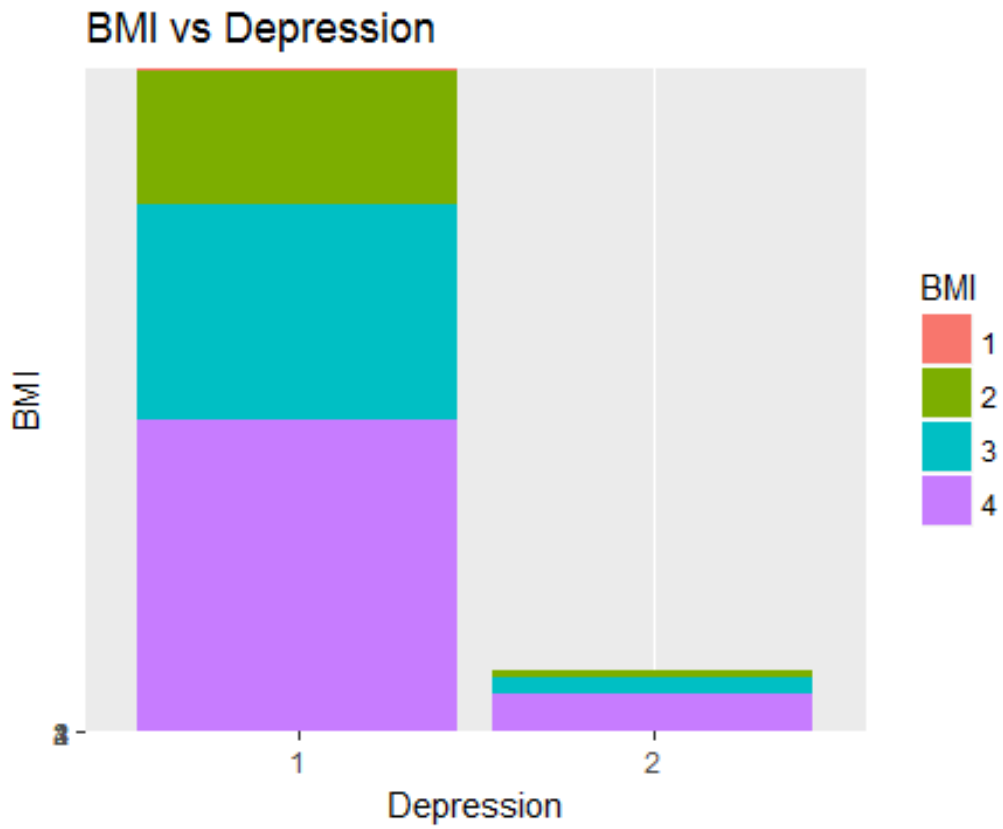


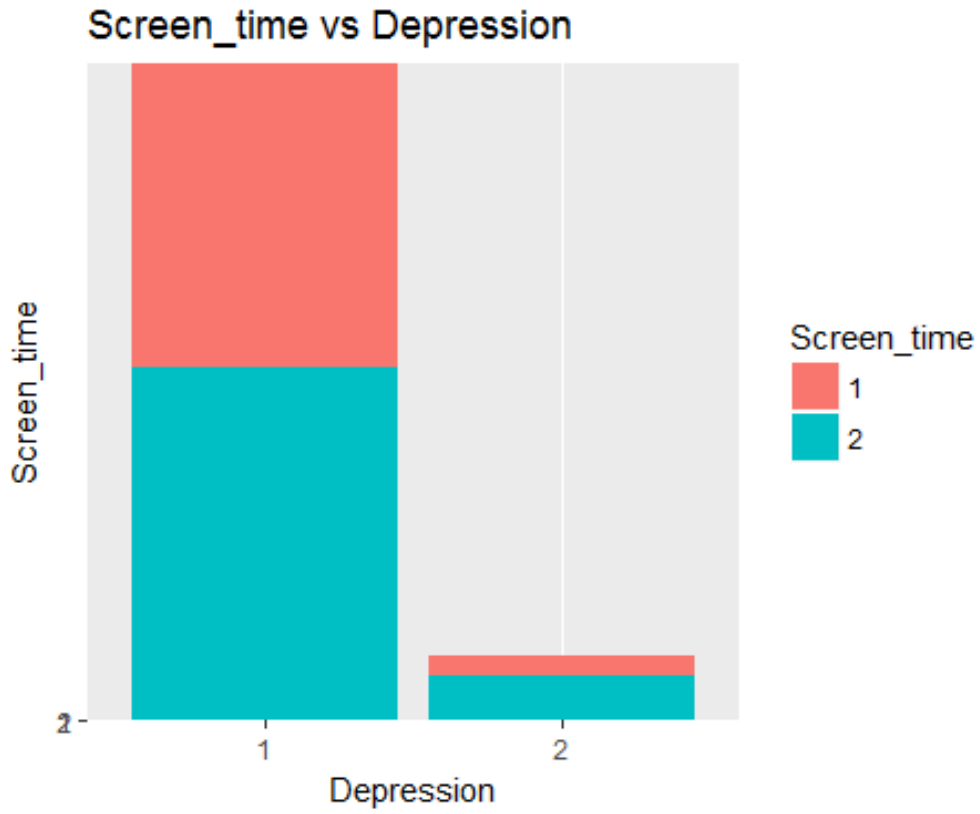

Age vs Depression
Age
1
2
3
4
Depression


Education vs Depression
Education
1
2
Depression


Race vs Depression
Race
1
2
3
4
1
2
Depression


Poverty vs Depression
Poverty
1
2
1
2
Depression


BMI vs Depression
BMI
1
2
3
4
1
2
Depression


Screen_time vs Depression
Screen_time
1
2
1
2
Depression

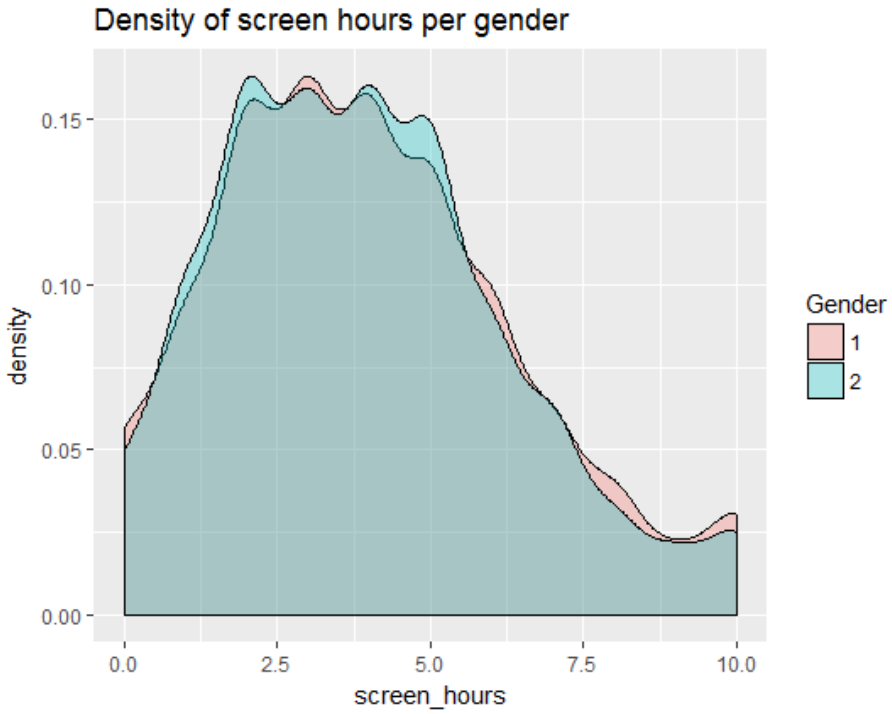
Comments on the above graphs

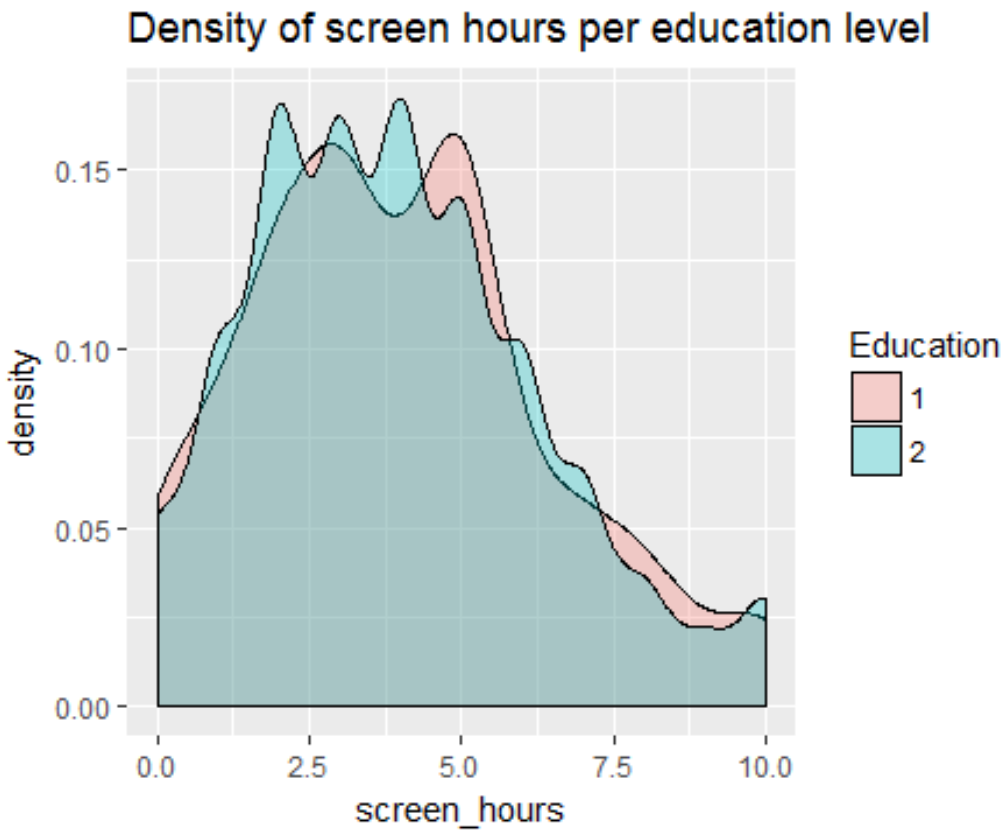
From the stacked bar plots, the following observations can be made about the association between each study variable and depression based on the sample:

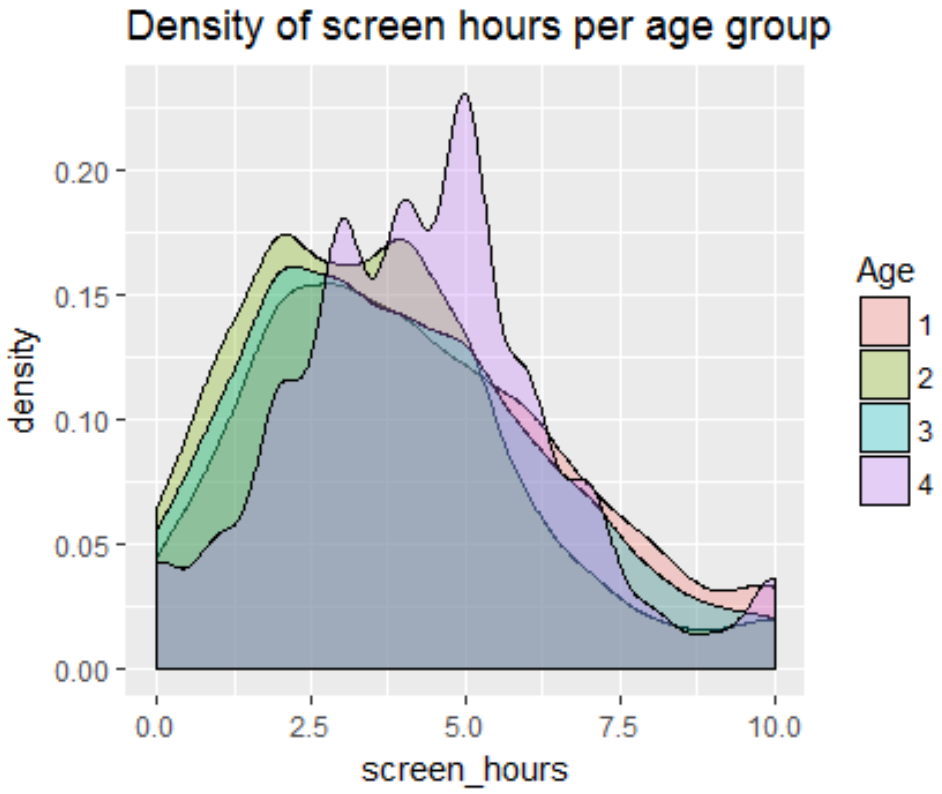
- Gender is an important factor for depression: women are much more likely to experience moderate or severe depression than men
- Screen time plays a significant role in depression: people who spent more time on TV/computer are much more likely to experience moderate or severe depression than people who spent less time on TV/computer

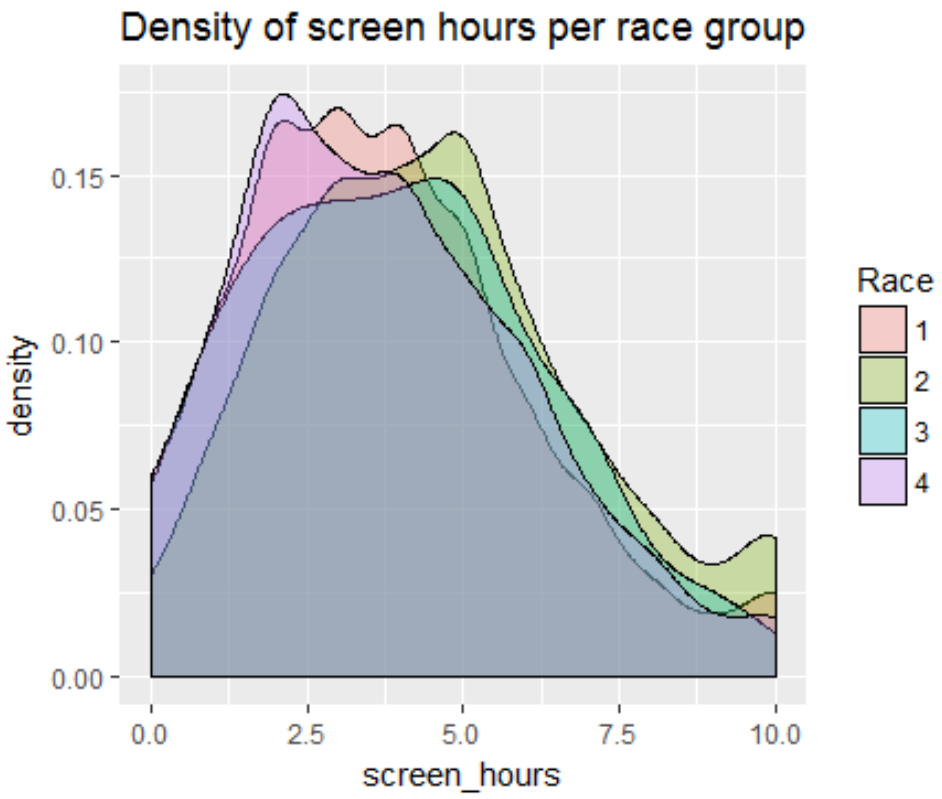
3.3 Association between screen hours and other study variables

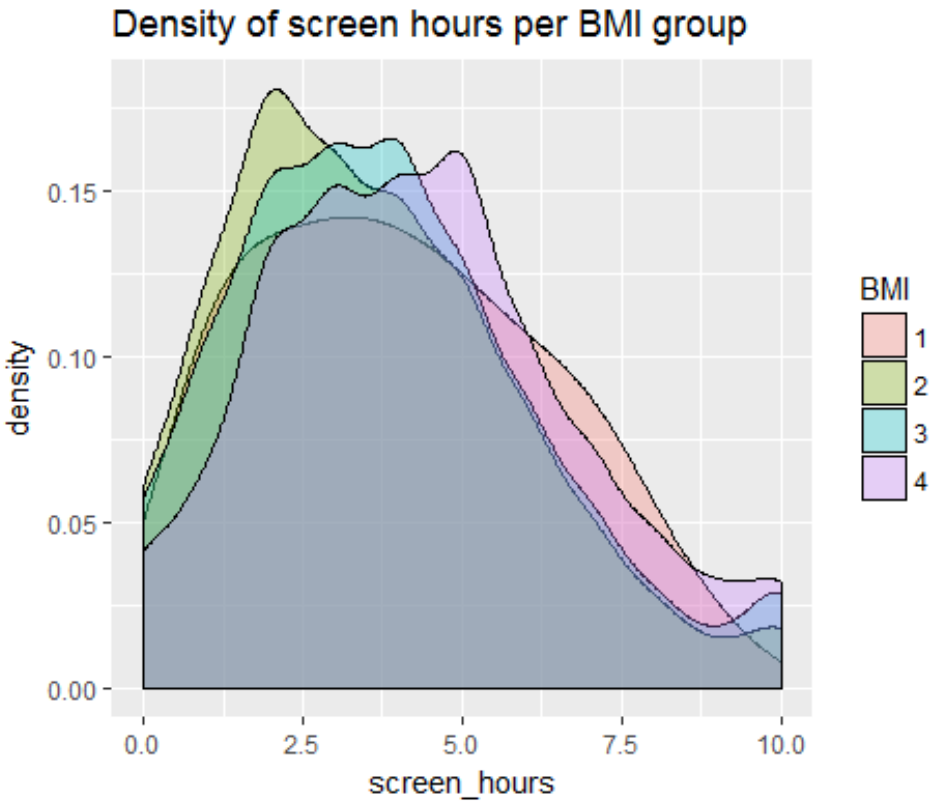
Since screen time is the main study variable in this study, it is worth to also examine the association between other study variables and screen time. The following kernel density plots are used to examine the association. Note that because kernel density plots work on continuous data, screen time here is represented by the total number of hours spent on TV and computers (which can be considered as a continuous variable in 0 to 10).

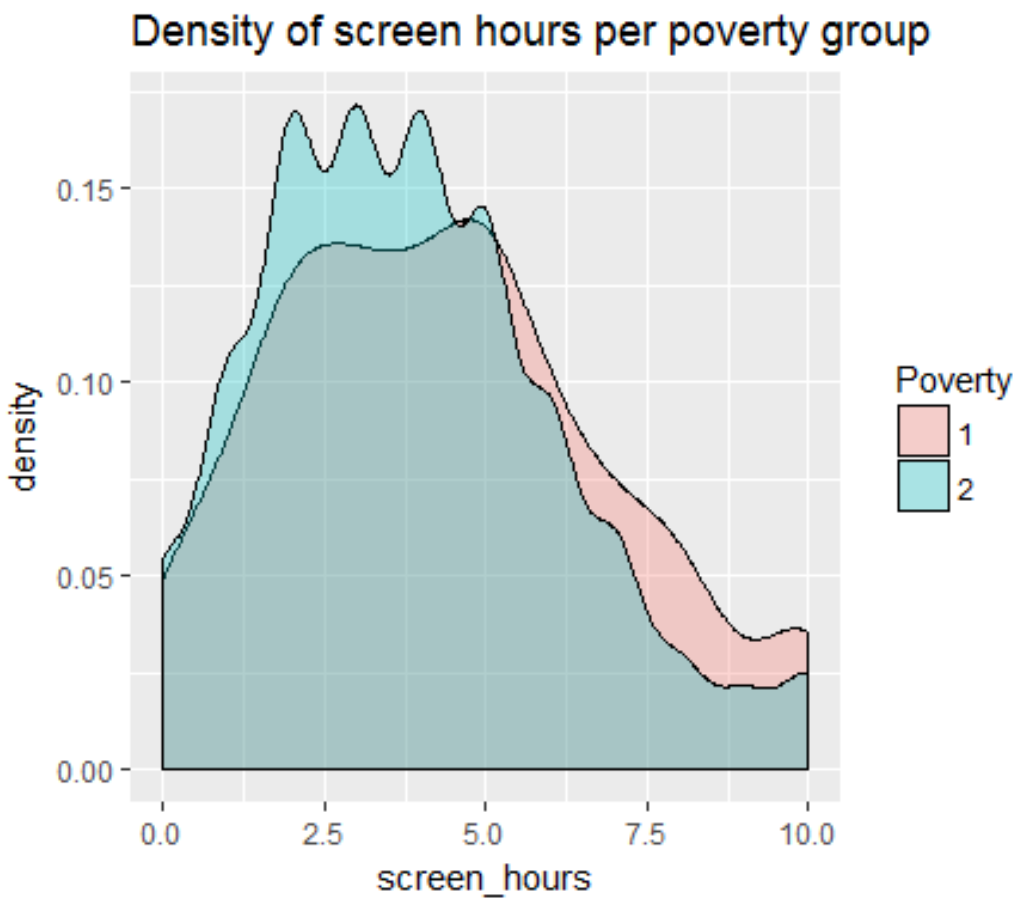



Density of screen hours per education level
density
0.15
0.10
0.05
0.00
0.0
2.5
5.0
7.5
10.0
screen_hours
Education
1
2


Density of screen hours per age group
density
0.20
0.15
0.10
0.05
0.00
0.0
2.5
5.0
7.5
10.0
screen_hours
Age
1
2
3
4


Density of screen hours per race group
density
0.15
0.10
0.05
0.00
0.0
2.5
5.0
7.5
10.0
screen_hours
Race
1
2
3
4


Density of screen hours per BMI group
density
0.15
0.10
0.05
0.00
0.0
2.5
5.0
7.5
10.0
screen_hours
BMI
1
2
3
4



Comments on the above graphs

From the kernel density plots above, it appears that there is no strong association between screen time and other study variables.

4. Estimation of population proportions using the sample data

4.1 Analysis 1: Weighted analysis

In order to estimate the population parameters using the sample, the following 3 additional variables in the sample are being used:

- SDMVPSU – Masked variance pseudo-PSU
- SDMVSTRA - Masked variance pseudo-Strata
- WTMEC2YR – Full sample 2 year MEC exam weight

By running weighted analysis using the survey package from R, descriptive statistics of the population are obtained and summarized in the following table as grouped by the study variables (n = 2,969):

Table 1.1 Descriptive statistics of the study population using weighted analysis

Study variable	**Sample size** n = 2,969	**No or mild depression (%)*** n = 2,731	**Moderate or severe depression (%)*** n = 238	**P-value from the Rao-Scott Chi-Square Test****
Gender				
Male	1,481	45.73 (0.88)	2.62 (0.41)	< 2.2 e-16
Female	1,488	46.89 (1.08)	4.77 (0.47)	
Education level				
Less than high school/GED	379	8.17 (1.25)	1.50 (0.28)	< 2.2 e-16
High school/GED	2,590	84.45 (1.56)	5.88 (0.69)	

equivalent or higher				
Race				
Non-Hispanic White	1,284	67.71 (2.71)	5.18 (0.76)	< 2.2 e-16
Non-Hispanic Black	725	8.79 (1.57)	0.83 (0.22)	
Hispanic	246	4.68 (0.99)	0.61 (0.24)	
Others	714	11.43 (1.52)	0.76 (0.17)	
Age				
20-35 years	1,035	30.24 (2.84)	2.40 (0.31)	< 2.2 e-16
36-50 years	798	26.09 (1.71)	2.44 (0.40)	
51-65 years	717	25.10 (1.12)	2.07 (0.26)	
> 65 years	419	11.18 (0.99)	0.48 (0.15)	
BMI				
Underweight	44	0.98 (0.22)	0.25 (0.06)	< 2.2 e-16
Normal	882	26.98 (2.17)	1.69 (0.38)	
Overweight	955	31.39 (1.18)	2.16 (0.49)	
Obese	1,088	33.25 (1.69)	3.29 (0.47)	

Poverty				
Yes (below poverty threshold)	608	11.26 (1.94)	2.41 (0.44)	< 2.2 e-16
No (above poverty threshold)	2,361	81.35 (2.19)	4.97 (0.54)	
Screen time				
≤ 4 hours per day (Low)	1,839	63.52 (1.75)	3.43 (0.42)	< 2.2 e-16
> 4 hours per day (High)	1,130	29.10 (1.33)	3.95 (0.54)	

*Estimated population proportions are expressed as percentages with standard errors included in the parentheses

**Rao-Scott chi-square test with second order corrections (which corrects both mean and variance) is used to test the association between depression and the study variables. The Rao-Scott chi-square test is a design-adjusted version of the Pearson chi-square test that uses differences between observed and expected frequencies of the data. Here, the null hypothesis for each of the test is no association between depression and the study variable.

Note: In the original research paper, although the screen time variable is first built as a binary variable (low vs high), analysis on the variable is based on 3 categories (< 4 hours, 4-6 hours, > 6 hours). For consistency purposes, in this study all analyses are done based on 2 categories (low screen time and high screen time).

Conclusions from the above table

The following conclusions can be made about the population from table 1.1:

- About 7.4% of the whole study population had moderate or severe depression (PHQ-9 score of 10 or more)
- Both genders appear equally in the sample. A significant higher proportion of females (4.8%) had experienced moderate or severe depression as compare to males (2.6%) in the study population
- People who spent more time on TV watching/computer use (> 4 hours per day) appear to have much higher proportion of experiencing moderate or severe depression compared to people who spent less time on that (≤ 4 hours per day) in the study population
- It appears that each study variable is statistically significant to depression (as indicated by the p-values) on a univariate basis

In addition, it is noted that the above conclusions agree with the conclusions from the original research paper (except that in the original research paper it also mentioned that a higher proportion of moderate or severe depression appeared in the Non-Hispanic white population which is not significant as seen in this study). Overall, the estimated population proportions using the weighted analysis mostly agree with the estimations in the original research paper (again except the Race section), and the discrepancies can be explained by the slightly different sample dataset.

4.2 Analysis 2. Unweighted analysis

For the unweighted analysis, equal sampling weights are being used.

By running the unweighted analysis using the survey package from R, descriptive statistics of the population are obtained and summarized in the following table as grouped by the study variables.

Table 1.2 Descriptive statistics of the study population using unweighted analysis

Study variable	**Sample size** n = 2,969	**No or mild depression (%)*** n = 2,731	**Moderate or severe depression (%)*** n = 238	**P-value from the Rao-Scott Chi-Square Test****
Gender				
Male	1,481	47.09 (0.90)	2.80 (0.39)	< 2.2 e-16
Female	1,488	44.90 (1.03)	5.22 (0.61)	
Education level				
Less than high school/GED	379	10.95 (0.98)	1.82 (0.33)	< 2.2 e-16
High school/GED equivalent or higher	2,590	81.04 (1.49)	6.20 (0.68)	
Race				
Non-Hispanic White	1,284	39.27 (3.25)	3.97 (0.67)	< 2.2 e-16
Non-Hispanic Black	725	22.50 (2.86)	1.92 (0.41)	
Hispanic	246	7.34 (1.69)	0.94 (0.39)	
Others	714	22.87 (2.53)	1.18 (0.24)	
Age				

20-35 years	1,035	32.17 (2.30)	2.69 (0.39)	< 2.2 e-16
36-50 years	798	24.32 (1.34)	2.56 (0.37)	
51-65 years	717	22.09 (1.06)	2.05 (0.28)	
> 65 years	419	13.41 (0.93)	0.71 (0.16)	
BMI				
Underweight	44	1.28 (0.29)	0.20 (0.06)	< 2.2 e-16
Normal	882	28.16 (2.06)	1.55 (0.27)	
Overweight	955	30.01 (0.98)	2.16 (0.36)	
Obese	1,088	32.54 (1.30)	4.11 (0.54)	
Poverty				
Yes (below poverty threshold)	608	17.31 (1.79)	3.17 (0.56)	< 2.2 e-16
No (above poverty threshold)	2,361	74.67 (2.18)	4.85 (0.53)	
Screen time				
≤ 4 hours per day	1,839	58.24 (1.67)	3.70 (0.44)	< 2.2 e-16
> 4 hours per day	1,130	33.75 (1.17)	4.31 (0.53)	

4.3 Weighted vs unweighted analysis

As shown in table 1.3 below, the estimates and their standard errors are significantly different between the weighted and unweighted analyses. This makes sense as the NHANES sampling design uses unequal probability selection where the sample weights are different to reflect the number of people in the population that they represent. In this case, using an unweighted analysis will be inappropriate as equal sample probability increases the bias of the estimates as well as decreases the precision of the estimates. For example, it can be seen that the largest difference between the estimates and their standard errors occur in the race categories, and this can be explained by the fact that most of the oversampling of the design is done based on the race characteristics of the individuals (as mentioned earlier in the report), therefore, assuming equal sample weight will totally ignore the oversampling impact and decrease the reliability and precision of the estimated population proportions.

Table 1.3 Compare the obtained descriptive statistics of the study population of weighted vs unweighted analysis. Results are expressed in relative difference in percentages with the unweighted results being the reference group.

Study variable	**Sample size** n = 2,969	**No or mild depression (%)*** n = 2,731	**Moderate or severe depression (%) **** n = 238	**P-value from the Rao-Scott Chi-Square Test****
Gender				
Male	Same	-2.89 (-2.22)	-6.43 (5.13)	Same
Female	Same	4.43 (4.85)	-8.62 (-22.95)	
Education level				
Less than high school/GED	Same	-25.39 (27.6)	-17.58 (-15.15)	Same

High school/GED equivalent or higher	Same	4.2 (4.70)	-5.16 (1.47)	
Race				
Non-Hispanic White	Same	-60.93 (-45.10)	30.48 (13.43)	Same
Non-Hispanic Black	Same	-36.24 (41.42)	-56.77 (-46.34)	
Hispanic	Same	-50.02 (-39.92)	-35.11 (-38.46)	
Others	Same	-6.00 (23.48)	-35.59 (-29.17)	
Age				
20-35 years	Same	13.63 (5.66)	-10.78 (-20.51)	Same
36-50 years	Same	-16.63 (6.45)	-4.69 (8.11)	
51-65 years	Same	-23.44 (-24.14)	0.98 (-7.14)	
> 65 years	Same	-4.19 (5.34)	-32.39 (-6.25)	
BMI				
Underweight	Same	2.18 (30.00)	25.00 (0.00)	Same
Normal	Same	-34.95 (8.38)	9.03 (40.74)	
Overweight	Same	8.95 (0.46)	0.00 (36.11)	

Obese	Same	-13.78 (13.68)	-19.95 (-12.96)	
Poverty				
Yes (below poverty threshold)	Same	-2.89 (-2.22)	-23.97 (-21.43)	Same
No (above poverty threshold)	Same	4.43 (4.85)	2.47 (1.89)	
Screen time				
≤ 4 hours per day	Same	-25.39 (27.6)	-7.84 (-4.55)	Same
> 4 hours per day	Same	4.2 (4.70)	-8.35 (1.89)	

5. Statistical analysis

This section focuses on analyzing whether there is a significant association between screen time of TV watching/computer use and depression for the study population using statistical approaches. To do this, both univariate and multivariate analyses are performed (which are also done in the original research paper). All statistical analyses in this section are based on the weighted analysis from the previous section using the survey package from R.

5.1 Association between screen time and depression based on univariate analyses

In this section, univariate logistic regression analyses are used to investigate the association between depression and each of the study variable. The tests focus on individuals who experienced moderate or severe depression.

Table 2.1 Unadjusted odds ratios and confidence intervals for moderate or severe depression population

Study variable	**Reference group**	**Unadjusted OR**	**95% CI**	**p-value**
Screen time				
> 4 hours (High)	≤ 4 hours (Low)	2.52	(1.79,3.54)	7.06e-05
Gender				
Female	Male	1.78	(1.40,2.26)	0.000278
Education level				
High school/GED equivalent or higher	Less than high school/GED	0.38	(0.24,0.60)	0.00087

BMI				
Underweight	Normal	4.02	(1.84,8.78)*	0.00363
Overweight		1.10	(0.56,2.18)	> 0.05**
Obese		1.59	(0.93,2.70)	> 0.05**
Poverty level				
No (above poverty threshold)	Yes (below poverty threshold)	0.29	(0.21,0.38)	3.59e-07

*The large standard error of the odds ratio indicates that the result may not be reliable to make any conclusion

**The large p-values indicate that the results may not be reliable to make any conclusion

Comments

The univariate logistic regression analyses have indicated the following results:

- Individuals who spent more than 4 hours per day on TV/computer are more likely to experience moderate or severe depression than people who spent less time on TV/computers (OR=2.52, 95% CI = 1.79-3.54)
- Females are more likely to experience moderate or severe depression than males (OR=1.78, 95% CI = 1.40-2.26)
- Individuals who have higher education are at reduced risk of experiencing moderate or severe depression than individuals who have lower education (OR=0.38, 95% CI = 0.24-0.60)
- Individuals who are above the poverty level are at reduced risk of experiencing moderate or severe depression than individuals who are in the poverty class (OR=0.29, 95% CI = 0.21-0.38)

Most of the above conclusions are also consistent with the conclusions in the original research paper (except that in this study the impact of age and race groups appears to be not significant while in the original paper some age and race groups are significant).

5.2 Association between screen time and depression based on multivariate analyses

In this section, multiple logistic regression analyses are used to investigate the association between screen time and depression, while treating all the other study variables (gender, education level, race, age, BMI, and poverty level) as confounding variables (i.e. keep them constant). By fitting the multiple logistic regression model, it is noticed that age, race, and BMI groups are insignificant statistically (with p-value > 0.05), which are consistent with the results in the univariate analysis. Therefore, the final selected model involves only the following variables: gender, education level, poverty level, and screen time. The interaction terms between screen time and other variables are not included as the other variables do not seem to have a strong association with screen time (indicated by the density plots in earlier section), although this is an assumption that still needs further analysis to be done. Table 2.2 below shows the odds ratios (along with 95% CIs) of experiencing moderate or severe depression for each of the variable against their reference group.

Table 2.2 Adjusted odds ratios and confidence intervals for moderate or severe depression population

Study variable	**Reference group**	**Adjusted OR**	**95% CI**	**p-value**
Screen time				
> 4 hours (High)	≤ 4 hours (Low)	2.26	(1.57,3.24)	0.000685
Gender				
Female	Male	1.82	(1.39,2.39)	0.000862

Education level				
High school/GED equivalent or higher	Less than high school/GED	0.48	(0.29,0.82)	0.018013
Poverty level				
No (above poverty threshold)	Yes (below poverty threshold)	0.35	(0.24,0.50)	9.26e-05

Comments

The multiple logistic regression analyses have indicated the following results:

- Individuals who spent more than 4 hours per day on TV/computer are more likely to experience moderate or severe depression than people who spent less time on TV/computers (OR=2.26, 95% CI = 1.57-3.24)
- Females are more likely to experience moderate or severe depression than males (OR=1.82, 95% CI = 1.39-2.39)
- Individuals who have higher education are at reduced risk of experiencing moderate or severe depression than individuals who have lower education (OR=0.48, 95% CI = 0.29-0.82)
- Individuals who are above the poverty level are at reduced risk of experiencing moderate or severe depression than individuals who are in the poverty class (OR=0.35, 95% CI = 0.24-0.50)

The results from the multiple logistic regression analyses are mostly consistent with the conclusions from the univariate analyses (with some slight difference) and also consistent with the multivariate analysis results in the original research paper.

6. Final conclusions and remarks

This study uses a large population-based cross-sectional survey, the 2011-2012 National Health and Nutrition Examination survey (NHANES), to investigate the association between the amounts of time people spent on TV watching and computer use with experiencing moderate or severe depression by using statistical methods on complex sample data. Similar to the conclusions from the original research paper that this study is based on, it is found that the odds of experiencing moderate or severe depression are much higher among individuals who spent more than 4 hours per day on TV watching and computer use outside of school or work. In addition, the study also indicates that females are more likely to experience moderate or severe depression than males in the study population. The conclusions in this study are based on both descriptive statistics of the study population as well as regression analysis of the data (both univariate and multivariate analysis). Today, as one of the leading cause of disease among North Americans, depression and mental health starts to bring more attention from the society and conclusions from this study suggests that people should reduce their time spent on TV watching and computer use outside of school or work in order to improve their mental health status.

Finally, there are a few limitations of this study. First, the sample dataset is heavily based on the self-reported responses from the participated individuals and this could introduce response/non-response bias to the results (for example, some people may report false income level, and some people with depression concerns may refuse to report their depression status and this could lead to underestimated depression rate). Second, some important risk factors are not being included in this study which may affect the multivariate analyses significantly. Further studies on this topic can focus on including more risk factors such as family history of depression, crime history, and sleep duration of the participated individuals, as well as performing more analysis on the response/non-response bias of the results in order to better examine the association between depression and individual behaviours.

Appendix (R code)

1) Derive the study variables and build the datasets using the 2011-2012 NHANES data

```
rm(list=ls(all=TRUE))

require(foreign)

setwd('D:/CHL7001/Final Project')
```

```
##Questionnaire data (PHQ-9) - Mental Health - Depression Screener (DPQ_G)

DPQ1112=read.xport("DPQ1112.xpt")

dim(DPQ1112) #dimension of the data

colnames(DPQ1112) #obtain the column names

str(DPQ1112) #check the class/data type of each column/feature

#Ignore the response of the 10th question

DPQ1112 = subset(DPQ1112,select=-DPQ100)

#Remove rows which have SEQN missing

DPQ1112 = DPQ1112[is.na(DPQ1112[,1]) != 1, ]

#If the response is refused or don't know, treat it as NA

DPQ1112[DPQ1112==7 | DPQ1112==9] <- NA

#Remove rows which have all NAs in the response

DPQ1112 = DPQ1112[rowSums(is.na(DPQ1112[,2:10])) != 9, ]

#PHQ9 score = sum of the scores of the first 9 questions

PHQ9 = rowSums(DPQ1112[2:10], na.rm = TRUE)

#Categorize each person into No or mild depression OR Moderate to severe depression

#based on the PHQ9 score
```

```
Depression = NULL
for (i in 1:length(PHQ9)){
  if (PHQ9[i] <= 9) {
    Depression[i] = 1 #No or mild
  }
  else {
    Depression[i] = 2 #Moderate to severe
  }
}
#The final dataset will only be composed of the Sequence Number and Depression status
DPQ = as.data.frame(cbind(DPQ1112[,1],Depression))
colnames(DPQ) = c('SEQN','Depression')
##############################################
##Demographics data - Demographic Variables & Sample Weights (DEMO_G)
DEMO1112=read.xport("DEMO1112.xpt")
dim(DEMO1112) #dimension of the data:9107 21
colnames(DEMO1112) #obtain the column names: important ones are SEQN,
RIDAGEYR(age),RIAGENDR(gender), DMDEDUC2(education
                #level of age20+), RIDRETH1(race), INDFMPIR(income-poverty ratio)
str(DEMO1112) #check the class/data type of each column/feature
#First, filter out individuals with age<20 years
DEMO1112 = DEMO1112[which(DEMO1112$RIDAGEYR>=20),]
#Select only the columns needed
```

```
DEMO1112 =
subset(DEMO1112,select=c(SEQN,RIDAGEYR,RIAGENDR,DMDEDUC2,RIDRETH1,INDF
MPIR))

#Create the age variable based on different categories

age = NULL

for (i in 1:dim(DEMO1112)[1]){

  if (DEMO1112$RIDAGEYR[i] <= 35) {

    age[i] = 1 #20-35 years

  }

  else if (DEMO1112$RIDAGEYR[i] <= 50) {

    age[i] = 2 #36-50 years

  }

  else if (DEMO1112$RIDAGEYR[i] <= 65) {

    age[i] = 3 #51-65 years

  }

  else if (DEMO1112$RIDAGEYR[i] > 65) {

    age[i] = 4 #>65 years

  }

  else {

    age[i] =  NA

  }

}

#Create the educationa variable
```

```
education = NULL
for (i in 1:dim(DEMO1112)[1]){
  if (DEMO1112$DMDEDUC2[i] == 1 | DEMO1112$DMDEDUC2[i] == 2) {
    education[i] = 1 #Less than high school/GED
  }
  else if (DEMO1112$DMDEDUC2[i] == 3 | DEMO1112$DMDEDUC2[i] == 4 |
DEMO1112$DMDEDUC2[i] == 5) {
    education[i] = 2 #High school graduate or more
  }
  else {
    education[i] = NA
  }
}
#Create the race variable
race = NULL
for (i in 1:dim(DEMO1112)[1]){
  if (DEMO1112$RIDRETH1[i] == 3) {
    race[i] = 1 #Non-Hispanic White
  }
  else if (DEMO1112$RIDRETH1[i] == 4) {
    race[i] = 2 #Non-Hispanic Black
  }
  else if (DEMO1112$RIDRETH1[i] == 2) {
```

```
    race[i] = 3 #Hispanic
  }
  else if (DEMO1112$RIDRETH1[i] == 1 | DEMO1112$RIDRETH1[i] == 5) {
    race[i] = 4 #Other race
  }
  else {
    race[i] = NA
  }
}
#Create the poverty variable
poverty = NULL
for (i in 1:dim(DEMO1112)[1]) {
  if (is.na(DEMO1112$INDFMPIR[i]) ==  TRUE) {
    poverty[i] = NA
  }
  else if (DEMO1112$INDFMPIR[i] < 1) {
    poverty[i] = 1 #Yes
  }
  else {
    poverty[i] = 2 #No
  }
}
#Combine the above datasets into a single dataset
```

```
DEMO =
as.data.frame(cbind(DEMO1112$SEQN,DEMO1112$RIAGENDR,age,education,race,poverty))
colnames(DEMO) =  c('SEQN','Gender','Age','Education','Race','Poverty')
#Filter out individuals without complete information (i.e. one or more NAs)
DEMO = DEMO[complete.cases(DEMO),]
##############################################
##Examination data - Body Measures (BMX_G)
BMX1112=read.xport("BMX1112.xpt") ### bmi
dim(BMX1112) #dimension of the data:9107 21
colnames(BMX1112) #obtain the column names: important ones are SEQN, PAQ710, PAQ715
str(BMX1112) #check the class/data type of each column/feature
#Select only the SEQN and BMXBMI columns
BMX1112 = subset(BMX1112,select=c(SEQN,BMXBMI))
#Exclude individuals who do not have complete responses
BMX1112 = BMX1112[complete.cases(BMX1112), ]
#Categorize each individual into 1 of the 4 BMI categories
BMI_Group = NULL
for (i in 1:dim(BMX1112)[1]){
  if (BMX1112[i,2] < 18.5) {
    BMI_Group[i] = 1 #Underweight
  }
  else if (BMX1112[i,2] < 24.9) {
    BMI_Group[i] = 2 #Normal weight
```

```
  }
  else if (BMX1112[i,2] < 29.9) {
    BMI_Group[i] = 3 #Overweight
  }
  else {
    BMI_Group[i] = 4 #Obese
  }
}
#The final dataset will only be composed of the Sequence Number and BMI_Group
BMI = as.data.frame(cbind(BMX1112$SEQN,BMI_Group))
colnames(BMI) = c('SEQN','BMI')
################################################
##Questionnaire data - Physical Activity (PAQ_G)
PAQ1112=read.xport("PAQ1112.xpt") ### tv watching/computer use
dim(PAQ1112) #dimension of the data:9107 21
colnames(PAQ1112) #obtain the column names: important ones are SEQN, PAQ710, PAQ715
str(PAQ1112) #check the class/data type of each column/feature
#Select only the SEQN, PAQ710, and PAQ715 columns
PAQ1112 = subset(PAQ1112,select=c(SEQN,PAQ710,PAQ715))
#If the response is don't watch TV/computer or refused or don't know, treat it as NA
PAQ1112[PAQ1112==8 | PAQ1112==77 | PAQ1112==99] <- NA
```

```
#Exclude individuals who do not have complete responses for the 2 questions (i.e. have one or more NAs)

PAQ1112 = PAQ1112[complete.cases(PAQ1112), ]

#calculate the screen score = sum of the scores of the 2 questions

screen_score = rowSums(PAQ1112[2:3])

#Categorize each person into Low screen time OR High screen time based on the screen score

screen_time = NULL

for (i in 1:length(screen_score)) {

  if (screen_score[i] <= 4) {

    screen_time[i] = 1 #Low

  }

  else {

    screen_time[i] = 2 #High

  }

}

#The final dataset will only be composed of the Sequence Number and Screen_time

PAQ = as.data.frame(cbind(PAQ1112$SEQN,screen_time))

colnames(PAQ) = c('SEQN','Screen_time')

##############################################

#Now, merge all datasets together by SEQN

dat1 = merge(DPQ,DEMO, by="SEQN")

dat2 = merge(dat1,BMI, by="SEQN")

dat = merge(dat2,PAQ, by="SEQN")
```

```
dim(dat) #2970 data, 8 variables

#Obtain the survey sampling weights for each of the individual

#of the selected sample dataset

DEMO1112=read.xport("DEMO1112.xpt")

weights =
subset(DEMO1112,select=c(SEQN,SDMVPSU,SDMVSTRA,WTINT2YR,WTMEC2YR))

#Append the sampling weights to the created dataset

dat = merge(dat,weights, by='SEQN')

#convert all 8 variables from numeric type to factor type

for (i in 2:11) {

  dat[,i] = as.factor(dat[,i])

}

str(dat)

save(dat,file="data.Rda") #save the merged dataset in the local drive
```

---

2) Exploratory data analysis using ggplot2

```
setwd('D:/CHL7001/Final Project')

load('data.rda')

library(ggplot2)
```

```
##Relationship between explanatory and response variable -Barplots showing the distribution of
#factors within each variable

colour_plate = c('skyblue','goldenrod','dodgerblue','coral',

            'chartreuse','brown','cyan','cornsilk')

ggplot(dat, aes(x=Depression)) +
```

```
    geom_bar(fill=colour_plate[1],colour='black') +
    ggtitle('Depression status')
ggplot(dat, aes(x=Age)) +
  geom_bar(fill=colour_plate[2],colour='black') +
  ggtitle('Age group')
ggplot(dat, aes(x=Gender)) +
  geom_bar(fill=colour_plate[3],colour='black') +
  ggtitle('Gender')
ggplot(dat, aes(x=Education)) +
  geom_bar(fill=colour_plate[4],colour='black') +
  ggtitle('Education group')
ggplot(dat, aes(x=Race)) +
  geom_bar(fill=colour_plate[5],colour='black') +
  ggtitle('Race category')
ggplot(dat, aes(x=Poverty)) +
  geom_bar(fill=colour_plate[6],colour='black') +
  ggtitle('Poverty')
ggplot(dat, aes(x=BMI)) +
  geom_bar(fill=colour_plate[7],colour='black') +
  ggtitle('BMI category')

ggplot(dat, aes(x=Screen_time)) +
  geom_bar(fill=colour_plate[8],colour='black') +
```

```
  ggtitle('Screen time')
##################################################
#Stacked barplots showing the association between each of the explanatory variable
#and the response variable (Depression):
ggplot(dat, aes(x=Depression,y=Gender,fill=Gender)) +
  geom_bar(stat='identity') +
  ggtitle('Gender vs Depression')
ggplot(dat, aes(x=Depression,y=Age,fill=Age)) +
  geom_bar(stat='identity') +
  ggtitle('Age vs Depression')
ggplot(dat, aes(x=Depression,y=Education,fill=Education)) +
  geom_bar(stat='identity') +
  ggtitle('Education vs Depression')
ggplot(dat, aes(x=Depression,y=Race,fill=Race)) +
  geom_bar(stat='identity') +
  ggtitle('Race vs Depression')
ggplot(dat, aes(x=Depression,y=Poverty,fill=Poverty)) +
  geom_bar(stat='identity') +
  ggtitle('Poverty vs Depression')
ggplot(dat, aes(x=Depression,y=BMI,fill=BMI)) +
  geom_bar(stat='identity') +
  ggtitle('BMI vs Depression')
ggplot(dat, aes(x=Depression,y=Screen_time,fill=Screen_time)) +
```

```
  geom_bar(stat='identity') +
  ggtitle('Screen_time vs Depression')
#####################################################
##Relationship between each variable and screen time
#create a second dataset that also includes total screen hours
screen_hours = as.data.frame(cbind(PAQ1112$SEQN,screen_score))
colnames(screen_hours) = c('SEQN','screen_hours')
dat2 =  merge(dat,screen_hours, by='SEQN')
ggplot(dat2, aes(x = screen_hours, fill = Gender)) +
  geom_density(alpha = .3) + ggtitle('Density of screen hours per gender')
ggplot(dat2, aes(x = screen_hours, fill = Education)) +
  geom_density(alpha = .3) + ggtitle('Density of screen hours per education level')
ggplot(dat2, aes(x = screen_hours, fill = Age)) +
  geom_density(alpha = .3) + ggtitle('Density of screen hours per age group')
ggplot(dat2, aes(x = screen_hours, fill = Race)) +
  geom_density(alpha = .3) + ggtitle('Density of screen hours per race group')
ggplot(dat2, aes(x = screen_hours, fill = BMI)) +
  geom_density(alpha = .3) + ggtitle('Density of screen hours per BMI group')
ggplot(dat2, aes(x = screen_hours, fill = Poverty)) +
  geom_density(alpha = .3) + ggtitle('Density of screen hours per poverty group')
```

3) Estimation of population parameters using the survey package

```
setwd('D:/CHL7001/Final Project')

load('data.rda')

require(foreign)

library(survey)

library(dplyr)
```

```
#Summarize the sample size grouped by study variables

dat %>% group_by(Depression) %>% count()

dat %>% group_by(Gender) %>% count()

dat %>% group_by(Education) %>% count()

dat %>% group_by(Race) %>% count()

dat %>% group_by(Age) %>% count()

dat %>% group_by(BMI) %>% count()

dat %>% group_by(Poverty) %>% count()

dat %>% group_by(Screen_time) %>% count()
```

```
##Model 1: Weighted analysis on the sampling design

#1) Build the svydesign object using the sampling design

design1 =
svydesign(id=~SDMVPSU,strata=~SDMVSTRA,weights=~WTMEC2YR,nest=TRUE,data=dat
) #weighted analysis

#2) Use the design object to estimate the population the based on

#the study variables as well as testing the association
```

```
#1. Gender

#estimate the parameters and make it into a nice-looking table format

var1 = svymean(~interaction(Depression,Gender), design=design1)

table1 = ftable(var1, rownames=list(Depression=c('No or mild','Moderate or severe'),

                    Gender=c("Male",'Female')))

round(100*table1,2)

#Use Rao-Scott Chi-square test to the association between depression and the study variable

svychisq(~interaction(Depression,Gender),design1,statistic="F")

#2. Education level

var2 = svymean(~interaction(Depression,Education), design=design1)

table2 = ftable(var2, rownames=list(Depression=c('No or mild','Moderate or
severe'),Education=c("Less than high school/GED",'High school/GED equivalent or higher')))

round(100*table2,2)

svychisq(~interaction(Depression,Education),design1,statistic="F")

#3. Race

var3 = svymean(~interaction(Depression,Race), design=design1)

table3 = ftable(var3, rownames=list(Depression=c('No or mild','Moderate or
severe'),Race=c("Non-Hispanic White",'Non-Hispanic Black','Hispanic','Others')))

round(100*table3,2)

svychisq(~interaction(Depression,Race),design1,statistic="F")

#4. Age

var4 = svymean(~interaction(Depression,Age), design=design1)
```

```
table4 = ftable(var4, rownames=list(Depression=c('No or mild','Moderate or severe'),Age=c("20-35 years",'36-50 years','51-65 years','> 65 years')))

round(100*table4,2)

svychisq(~interaction(Depression,Age),design1,statistic="F")

#5. BMI

var5 = svymean(~interaction(Depression,BMI), design=design1)

table5 = ftable(var5, rownames=list(Depression=c('No or mild','Moderate or severe'),BMI=c("Underweight",'Normal','Overweight','Obese')))

round(100*table5,2)

svychisq(~interaction(Depression,BMI),design1,statistic="F")

#6. Poverty

var6 = svymean(~interaction(Depression,Poverty), design=design1)

table6 = ftable(var6, rownames=list(Depression=c('No or mild','Moderate or severe'),Poverty=c('Yes','No')))

round(100*table6,2)

svychisq(~interaction(Depression,Poverty),design1,statistic="F")

#7. Screen_time

var7 = svymean(~interaction(Depression,Screen_time), design=design1)

table7 = ftable(var7, rownames=list(Depression=c('No or mild','Moderate or severe'),Poverty=c('<=4hours','>4hours')))

round(100*table7,2)

svychisq(~interaction(Depression,Screen_time),design1,statistic="F")

##############################################################

##Model 2: unWeighted analysis on the sampling design
```

```
#1) Build the svydesign object using the sampling design

design2 = svydesign(id=~SDMVPSU,strata=~SDMVSTRA,nest=TRUE,data=dat) #unweighted analysis: assuming equal probability

#2) Use the design object to estimate the population the based on

#the study variables as well as testing the association


#1. Gender

var1 = svymean(~interaction(Depression,Gender), design=design2)

table1 = ftable(var1, rownames=list(Depression=c('No or mild','Moderate or severe'),

                                    Gender=c("Male",'Female')))

round(100*table1,2)

svychisq(~interaction(Depression,Gender),design2,statistic="F")

#2. Education level

var2 = svymean(~interaction(Depression,Education), design=design2)

table2 = ftable(var2, rownames=list(Depression=c('No or mild','Moderate or severe'),Education=c("Less than high school/GED",'High school/GED equivalent or higher')))

round(100*table2,2)

svychisq(~interaction(Depression,Education),design2,statistic="F")

#3. Race

var3 = svymean(~interaction(Depression,Race), design=design2)

table3 = ftable(var3, rownames=list(Depression=c('No or mild','Moderate or severe'),Race=c("Non-Hispanic White",'Non-Hispanic Black','Hispanic','Others')))

round(100*table3,2)
```

```
svychisq(~interaction(Depression,Race),design2,statistic="F")

#4. Age

var4 = svymean(~interaction(Depression,Age), design=design2)

table4 = ftable(var4, rownames=list(Depression=c('No or mild','Moderate or severe'),Age=c("20-35 years",'36-50 years','51-65 years','> 65 years')))

round(100*table4,2)

svychisq(~interaction(Depression,Age),design2,statistic="F")

#5. BMI

var5 = svymean(~interaction(Depression,BMI), design=design2)

table5 = ftable(var5, rownames=list(Depression=c('No or mild','Moderate or severe'),BMI=c("Underweight",'Normal','Overweight','Obese')))

round(100*table5,2)

svychisq(~interaction(Depression,BMI),design2,statistic="F")

#6. Poverty

var6 = svymean(~interaction(Depression,Poverty), design=design2)

table6 = ftable(var6, rownames=list(Depression=c('No or mild','Moderate or severe'),Poverty=c('Yes','No')))

round(100*table6,2)

svychisq(~interaction(Depression,Poverty),design2,statistic="F")

#7. Screen_time

var7 = svymean(~interaction(Depression,Screen_time), design=design2)

table7 = ftable(var7, rownames=list(Depression=c('No or mild','Moderate or severe'),Poverty=c('<=4hours','>4hours')))

round(100*table7,2)
```

```
svychisq(~interaction(Depression,Screen_time),design2,statistic="F")
```

---

4) Use logistic regression to find the odds ratios and confidence intervals using the weighted analysis

```
setwd('D:/CHL7001/Final Project')

load('data.rda')

require(foreign)

library(survey)

library(dplyr)
```

```
##Univariate analysis

#1. Screen_time

model11 = svyglm(as.numeric(Depression==2) ~

                 Screen_time, design = design1, family=quasibinomial())

summary(model11)

#Calculate the unadjusted OR

m11 = model11$coefficients[2]

se11 = 0.1739

OR11 = exp(m11)

#Construct the 95% CI for the unadjusted OR

cat("  95% C.I.:  (", exp(m11-1.96*se11), ",", exp(m11+1.96*se11), ")\n", sep='')

#2. Age

model12 = svyglm(as.numeric(Depression==2) ~

                 Age, design = design1, family=quasibinomial())
```

```
summary(model12)

#Noted that age is not significant

#3. Gender

model13 = svyglm(as.numeric(Depression==2) ~

            Gender, design = design1, family=quasibinomial())

summary(model13)

#Calculate the unadjusted OR

m13 = model13$coefficients[2]

se13 = 0.1240

OR13 = exp(m13)

#Construct the 95% CI for the unadjusted OR

cat("  95% C.I.:  (", exp(m13-1.96*se13), ",", exp(m13+1.96*se13), ")\n", sep='')

#4. Education level

model14 = svyglm(as.numeric(Depression==2) ~

            Education, design = design1, family=quasibinomial())

summary(model14)

#Calculate the unadjusted OR

m14 = model14$coefficients[2]

se14 = 0.2380

OR14 = exp(m14)

#Construct the 95% CI for the unadjusted OR

cat("  95% C.I.:  (", exp(m14-1.96*se14), ",", exp(m14+1.96*se14), ")\n", sep='')

#5. Poverty level
```

```
model15 = svyglm(as.numeric(Depression==2) ~
                 Poverty, design = design1, family=quasibinomial())
summary(model15)
#Calculate the unadjusted OR
m15 = model15$coefficients[2]
se15 = 0.1514
OR15 = exp(m15)
#Construct the 95% CI for the unadjusted OR
cat("  95% C.I.:  (", exp(m15-1.96*se15), ",", exp(m15+1.96*se15), ")\n", sep='')
#6. BMI
#Note: for BMI, need to modify the dataset so that normal weight
#represents the reference group
BMI2 = NULL
BMI1 = as.numeric(dat$BMI)
dat2 = dat
for (i in 1:length(BMI1)) {
  if (BMI1[i] == 1) {
    BMI2[i] = 2
  }
  else if (BMI1[i] == 2) {
    BMI2[i] = 1
  }
  else {
```

```
   BMI2[i] = BMI1[i]
  }
}
dat2$BMI = as.factor(BMI2)
dat2 %>% group_by(BMI) %>% count() #check the modification
design11 =
svydesign(id=~SDMVPSU,strata=~SDMVSTRA,weights=~WTMEC2YR,nest=TRUE,data=dat
2)
model16 = svyglm(as.numeric(Depression==2) ~
                 BMI, design = design11, family=quasibinomial())
summary(model16)
#Unadjusted OR for Underweight group
m161 = model16$coefficients[2]
se161 = 0.39877
OR161 = exp(m161)
OR161
#Construct the 95% CI for the unadjusted OR
cat("  95% C.I.:  (", exp(m161-1.96*se161), ",", exp(m161+1.96*se161), ")\n", sep='')
#Noted that only the underweight group is significant
```

```
##Multivariate analysis
#Build the full model
model1 = svyglm(as.numeric(Depression==2)~
                  Gender+Education+Age+Race+
```

```
        BMI+Poverty+Screen_time, design = design1, family=quasibinomial())
summary(model1)
#Build the reduced model
model2 = svyglm(as.numeric(Depression==2)~
        Gender+Education+Poverty+Screen_time, design = design1, family=quasibinomial())
summary(model2)

#1. Gender
#Calculate the adjusted OR
m21 = model2$coefficients[2]
se21 = 0.1395
OR21 = exp(m21)
#Construct the 95% CI for the adjusted OR
cat("  95% C.I.:  (", exp(m21-1.96*se21), ",", exp(m21+1.96*se21), ")\n", sep='')
#2. Education level
#Calculate the adjusted OR
m22 = model2$coefficients[3]
se22 = 0.2677
OR22 = exp(m22)
OR22
#Construct the 95% CI for the adjusted OR
cat("  95% C.I.:  (", exp(m22-1.96*se22), ",", exp(m22+1.96*se22), ")\n", sep='')
```

```
#3. Poverty level

#Calculate the adjusted OR

m23 = model2$coefficients[4]

se23 = 0.1903

OR23 = exp(m23)

#Construct the 95% CI for the adjusted OR

cat("  95% C.I.:  (", exp(m23-1.96*se23), ",", exp(m23+1.96*se23), ")\n", sep='')

#4. Screen_time

#Calculate the adjusted OR

m24 = model2$coefficients[5]

se24 = 0.1840

OR24 = exp(m24)

#Construct the 95% CI for the adjusted OR

cat("  95% C.I.:  (", exp(m24-1.96*se24), ",", exp(m24+1.96*se24), ")\n", sep='')
```

References

Madhav, K., Sherchand, S. P., & Sherchan, S. (2017). Association between screen time and depression among US adults. *Preventive Medicine Reports, 8*, 67-71

Rothwell, C. J., & Madans, J. H. (2014). National Health and Nutrition Examination Survey: Sample Design, 2011-2014. *Vital and Health Statistics, 162,* 2nd ser. Retrieved July 29, 2018.

Murray, J. S. (n.d.). Intro to the survey R package (36-303). Retrieved July 29, 2018, from http://www.andrew.cmu.edu/user/jsmurray/teaching/303/files/lab.html

Lumley, T. (2007, March 16). Complex survey samples in R. Retrieved July 29, 2018, from http://r-survey.r-forge.r-project.org/survey/survey-wss.pdf

Scott, A. (n.d.). Rao-Scott corrections and their impact. *Section on Survey Research Methods,* 3514-3518. Retrieved July 29, 2018