FACTORS THAT AFFECT MENTAL HEALTH

By: Eric Brown

Table of Contents

Project Proposal

The data I will be conducting my ANCOVA analysis will be on mental health. Every year adults deal with mental health issues that sometimes go unnoticed by trained practitioners in the industry. Mainly because most people are still afraid to express their mental pain, openly to the public. Hopefully, the topic of mental health becomes less taboo, as we progress as a society. So, we can get a better understanding of what causes people to go through depression and/or experience anxiety or panic attacks.

However, for my analysis, luckily I was able to find data from CDC's website pertaining to mental health. More specifically, I was able to use a program called "The National Health and Nutrition Examination Survey" (NHANES) to get my data from, to conduct my analysis. Since NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations into meaningful data. NHANES has been part of the CDC for a while, providing vital health statistics for the Nation.

From NHANES's collection of data, I will try to test whether certain depression indicator variables are more statistically significant at causing people to experience depression every day of the year. Since NHANES only conducts its surveys on a yearly basis, my dataset will not be as long as it could be, if I were conducting my analysis on a monthly basis. But I should still be able to see which indicator variables stand out more, based on their means between two independent groups. Lastly, the ANCOVA test I will conduct, will be either one-way or two-way.

Model:

Depression ~ LevelofDepression + TroubleSleeping/SleepingtoMuch + Year_TwoYearInterval+
TroubleSleeping_OrsleepingTooMuch + FeelingTired_OrhavingLittleEnergy +
Have_LittleInterestDoingThings + Feelingbad_aboutYourself +
Level_ofDepression1_3:Year_TwoYearInterval

Let's go over the formula created above:

We will fit an ANCOVA model using the number of people who had a form of depression, as the response variable. Then we will have a total of eight main effects and one interaction term between two of the main effects in the model. The main effects will include the predictor variable of the mean level of depression faced by people in the sample population, ranking from 0-lowest level depression to 3-highest level of depression.

Then, the covariate depression indicator variables will include the count of people

1.) who had trouble sleeping or sleeping too much,
2.) who felt too tired or had little energy throughout the day
3.) who had trouble concentrating throughout the day
4.) who had little interest doing things throughout the day
5.) who felt bad about themselves
6.) who had suicidal thoughts and thoughts of being better off dead

The last main effect variable expressed in the model will be the date variable (year). Which will be based on the principle that each year would also include the preceding year's depression variable statistically count. For example, year 1999 would also include year 2000 depression indicator variable's data. That is why the year variable goes in sequences of every other year instead of sequential years. Lastly, the interaction term will be between the year variable and the factor predictor variable of the mean level of depression faced among the sample population.

Dataset Exploration

First, we should always explore our data. Before we begin creating our ANCOVA model in R, evaluating the significance of depression indicator variable. Just to make sure there are no outlier data or missing data values that could skew our results.

```r
#need this library to use read_excel import function
library(readxl)
#importing dataset from excel workbook
finalProject_data <- read_excel("C:/Users/esbro/Desktop/STAT 485/Project/FinalProject_data.xlsx",
                              col_types = c("numeric", "numeric", "numeric",
                                            "numeric", "numeric", "numeric",
                                            "numeric", "numeric", "text", "text"))
#need to factor both the Level of Depression & Covid Year
#since they values no do not represent count values
finalProject_data$Level_ofDepression1_3<-as.factor(finalProject_data$Level_ofDepression1_3)#scale variable 1-3
finalProject_data$Covid_year<-as.factor(finalProject_data$Covid_year) #binary variable 0 or 1
#getting summary of data values in the project dataset
#checking to make sure data values are consistent and not skewed
summary(finalProject_data)
```

```
##  Year_TwoYearInterval Depressed_Yes    TroubleSleeping_OrsleepingTooMuch
##  Min.   :1999         Min.   : 116.0   Min.   :115.0
##  1st Qu.:2004         1st Qu.: 452.0   1st Qu.:220.5
##  Median :2009         Median : 831.0   Median :429.0
##  Mean   :2009         Mean   : 723.4   Mean   :390.9
##  3rd Qu.:2014         3rd Qu.: 930.0   3rd Qu.:513.5
##  Max.   :2019         Max.   :1456.0   Max.   :764.0
##  FeelingTired_OrhavingLittleEnergy Trouble_concentratingOnThings
##  Min.   : 68.0                     Min.   : 77.0
##  1st Qu.:199.5                     1st Qu.: 93.5
##  Median :432.0                     Median :159.0
##  Mean   :365.0                     Mean   :151.1
##  3rd Qu.:493.5                     3rd Qu.:179.5
##  Max.   :697.0                     Max.   :291.0
##  Have_LittleInterestDoingThings Feelingbad_aboutYourself
##  Min.   : 56.0                  Min.   : 48.0
##  1st Qu.: 94.0                  1st Qu.: 70.0
##  Median :198.0                  Median :136.0
##  Mean   :170.5                  Mean   :122.4
##  3rd Qu.:230.0                  3rd Qu.:159.5
##  Max.   :310.0                  Max.   :207.0
##  Thoughts_youWould_betterOffDead Level_ofDepression1_3 Covid_year
##  Min.   :13.00                   1:6                   0:10
##  1st Qu.:31.00                   2:5                   1: 1
##  Median :33.00
##  Mean   :31.36
##  3rd Qu.:36.00
##  Max.   :40.00
```

```r
#checking the structure of the dataset, including variable types
str(finalProject_data)
```

```
## tibble [11 x 10] (S3: tbl_df/tbl/data.frame)
##  $ Year_TwoYearInterval             : num [1:11] 1999 2001 2003 2005 2007 ...
##  $ Depressed_Yes                    : num [1:11] 116 131 135 769 970 957 817 903 872 831 ...
##  $ TroubleSleeping_OrsleepingTooMuch: num [1:11] 119 121 115 320 526 543 429 501 419 443 ...
##  $ FeelingTired_OrhavingLittleEnergy: num [1:11] 68 69 73 326 525 495 400 492 438 432 ...
##  $ Trouble_concentratingOnThings    : num [1:11] 77 80 88 99 153 187 159 189 167 172 ...
##  $ Have_LittleInterestDoingThings   : num [1:11] 64 56 58 124 163 221 201 242 239 198 ...
##  $ Feelingbad_aboutYourself         : num [1:11] 51 51 48 89 156 163 154 167 136 124 ...
##  $ Thoughts_youWould_betterOffDead  : num [1:11] 33 32 36 13 40 33 36 32 30 24 ...
##  $ Level_ofDepression1_3            : Factor w/ 2 levels "1","2": 1 1 1 1 2 2 1 2 2 1 ...
##  $ Covid_year                       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

```r
#making sure no data values are missing before creating our ANCOVA model
sum(is.na(finalProject_data))
```

```
## [1] 0
```

```r
#install.packages("rstatix", repos = "https://cloud.r-project.org")
#rstatix provides a pipe-friendly framework, coherent with the 'tidyverse' design philosophy, for performing basi
c statistical tests
library(rstatix)
```

```
## Warning: package 'rstatix' was built under R version 4.1.3
```

```
##
## Attaching package: 'rstatix'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
# summary statistics for dependent variable depression based on the level of depression faced by people each year
finalProject_data %>% group_by(Level_ofDepression1_3) %>%
  get_summary_stats(Depressed_Yes, type="common")
```

```
## # A tibble: 2 x 11
##   Level_ofDepressio~ variable       n   min   max median   iqr  mean    sd    se
##   <fct>              <chr>      <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 1                  Depressed~     6   116   831    452   673  466.  372.  152.
## 2 2                  Depressed~     5   872  1456    957    67 1032.  241.  108.
## # ... with 1 more variable: ci <dbl>
```

```
finalProject_data %>% group_by(Covid_year) %>%
  get_summary_stats(Trouble_concentratingOnThings, type="common")
```

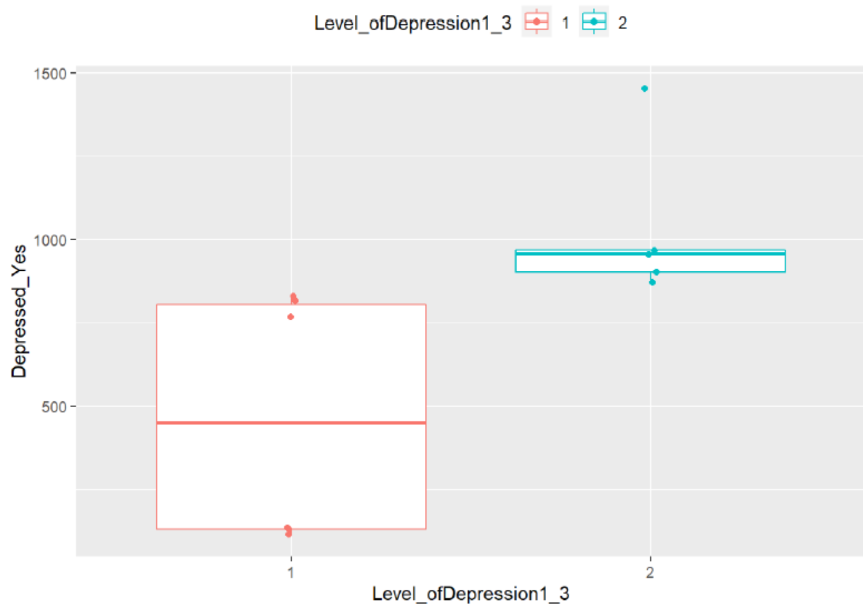
```
## Warning in stats::qt(alpha/2, .data$n - 1): NaNs produced
```

```
## # A tibble: 2 x 11
##   Covid_year variable         n   min   max median   iqr  mean    sd    se    ci
##   <fct>      <chr>        <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 0          Trouble_con~    10    77   189    156    80  137.  45.7  14.4  32.7
## 2 1          Trouble_con~     1   291   291    291     0  291    NA    NA   NaN
```

```
#more summary statistics involving factor variable level of depression faced among people each year
library(ggplot2)
ggplot(finalProject_data, aes(x = Level_ofDepression1_3,
                                        y = Depressed_Yes, col = Level_ofDepression1_3)) +
  geom_boxplot(outlier.shape = NA) + geom_jitter(width = 0.02) + theme(legend.position="top")
```

From the summary function, we can see that the data seems to be normally balanced, however we could further confirm this through a residual plot. Though there are some depression indicator variables that either have low data values or large ranges of values. These indicator variables include “thoughts of being better off dead” and “having little to no interest doing things.”


```
#more summary statistics involving factor variable level of depression faced among people each year
library(ggplot2)
ggplot(finalProject_data, aes(x = Level_ofDepression1_3,
                                        y = Depressed_Yes, col = Level_ofDepression1_3)) +
  geom_boxplot(outlier.shape = NA) + geom_jitter(width = 0.02) + theme(legend.position="top")
```



The boxplot also shows there was a wide range of values of people who expressed their depression levels a rate of 1 out of 3 each year. Compared to people who expressed their depression levels a rate of 2 out of 3 each year. Which makes sense because there were more people who had only a slight form of depression each year, than people who faced depression symptoms on a more extreme basis.

The covid year data grouping also influenced the frequency count of people who faced depression symptoms. Since we can see from the summary statistics involving one of the covariate variables "trouble concentrating throughout the day," the mean frequency of people who faced this depression symptom was less in years not during the coronavirus pandemic.

ANCOVA Model Assumptions:

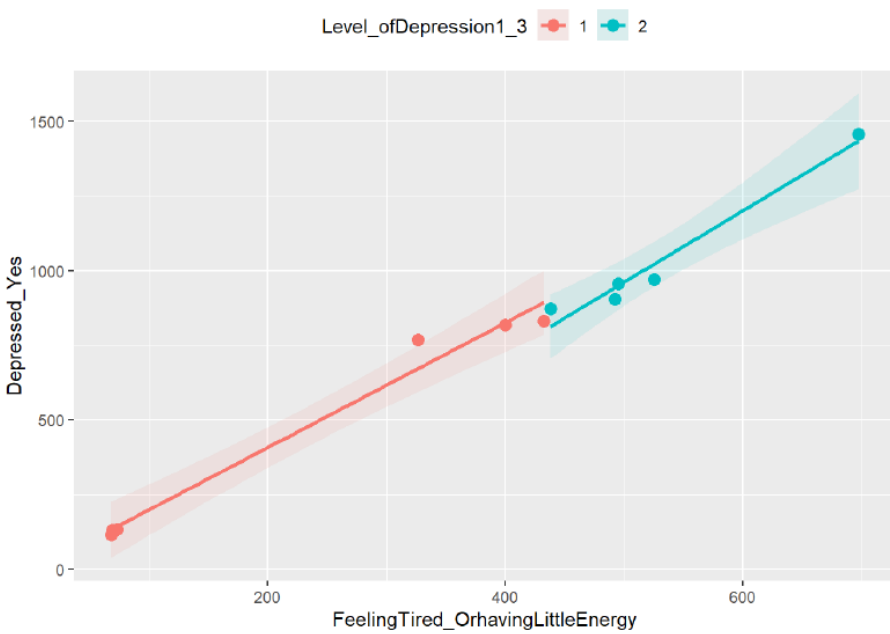
1.) *The relationship between the covariate variables and each group of the independent variable should be linear.*

```
#Checking to make sure ANCOVA assumptions are met
#linearity assumption
#the relationship between the covariates and at each group of the level of depression variable should be linear
ggplot(finalProject_data, aes(TroubleSleeping_OrsleepingTooMuch,
                              Depressed_Yes, colour = Level_ofDepression1_3)) + geom_point(size = 3) +
  geom_smooth(method = "lm", aes(fill = Level_ofDepression1_3), alpha = 0.1) + theme(legend.position="top")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

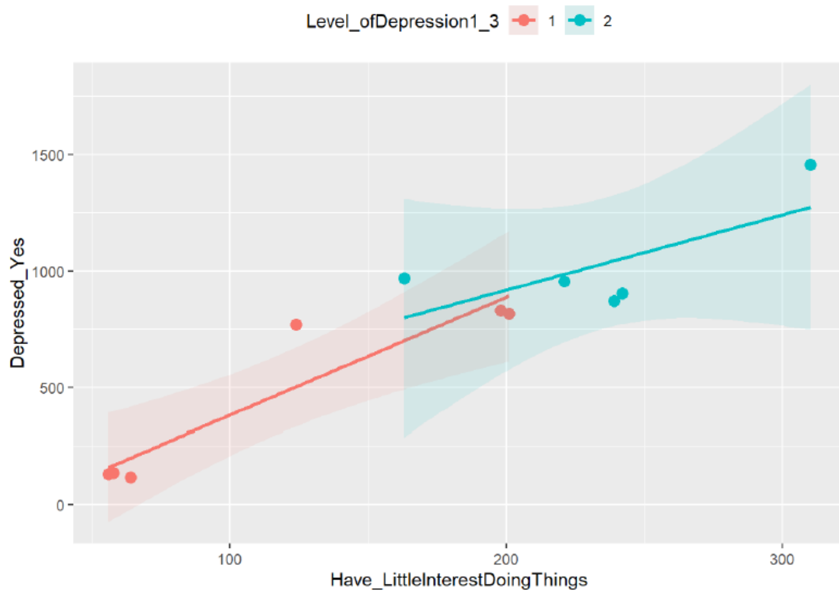

```
ggplot(finalProject_data, aes(FeelingTired_OrhavingLittleEnergy,
                              Depressed_Yes, colour = Level_ofDepression1_3)) + geom_point(size = 3) +
  geom_smooth(method = "lm", aes(fill = Level_ofDepression1_3), alpha = 0.1) + theme(legend.position="top")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



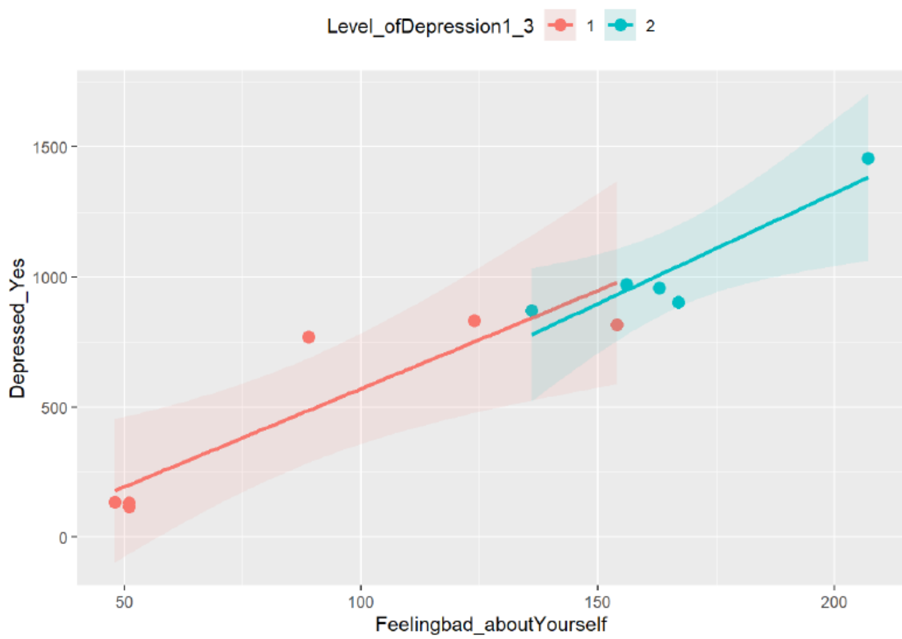

```
ggplot(finalProject_data, aes(Have_LittleInterestDoingThings,
                              Depressed_Yes, colour = Level_ofDepression1_3)) + geom_point(size = 3) +
  geom_smooth(method = "lm", aes(fill = Level_ofDepression1_3), alpha = 0.1) + theme(legend.position="top")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



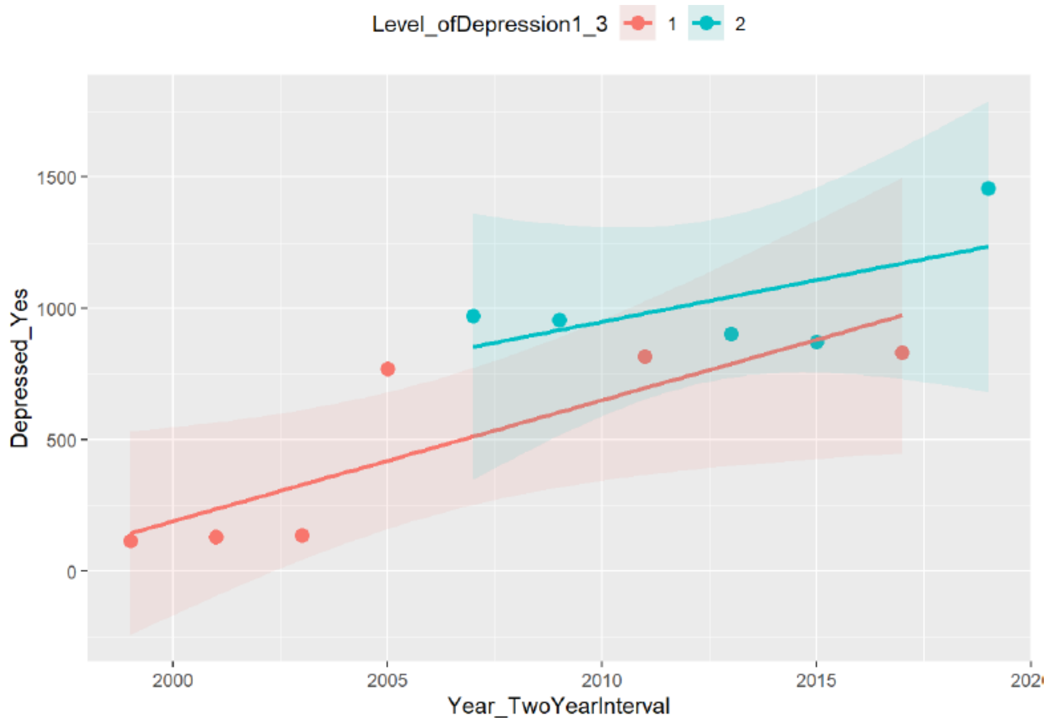

```
ggplot(finalProject_data, aes(Feelingbad_aboutYourself,
                              Depressed_Yes, colour = Level_ofDepression1_3)) + geom_point(size = 3) +
  geom_smooth(method = "lm", aes(fill = Level_ofDepression1_3), alpha = 0.1) + theme(legend.position="top")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#this scatterplot shows were show assumptions of linearity between the covariate and at each group of depression
 level
#might not be met, but it is mainly because of the outlier covid year of 2020
#we have an interaction term with this variable in the model to cover this issue
ggplot(finalProject_data, aes(Year_TwoYearInterval,
                              Depressed_Yes, colour = Level_ofDepression1_3)) + geom_point(size = 3) +
  geom_smooth(method = "lm", aes(fill = Level_ofDepression1_3), alpha = 0.1) + theme(legend.position="top")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



As we can see, for the most part the data points are closely in line with the fitted regression line. So we can safely say, linearity is met between covariates among each group of the independent variable "level of depression."

2.) *There should be no interaction between the categorical independent variable and covariate variables*

```
#our depression model to be used for analysis
depression.model <- lm(Depressed_Yes~ Level_ofDepression1_3 + Year_TwoYearInterval+
                         TroubleSleeping_OrsleepingTooMuch + FeelingTired_OrhavingLittleEnergy+
                         Have_LittleInterestDoingThings + Feelingbad_aboutYourself+
                         Level_ofDepression1_3:Year_TwoYearInterval, data = finalProject_data)


#testing assumption of no interaction between the categorical independent variable and covariate
#also finding the significance of interaction term in the model  with all covariates
Anova(aov(depression.model, data = finalProject_data), type = 3)
```

```
#isolating the interaction term to make sure it still is not significant by itself against response variable
Anova(aov(Depressed_Yes~Level_ofDepression1_3*Year_TwoYearInterval,
          data =finalProject_data), type = 3)
```

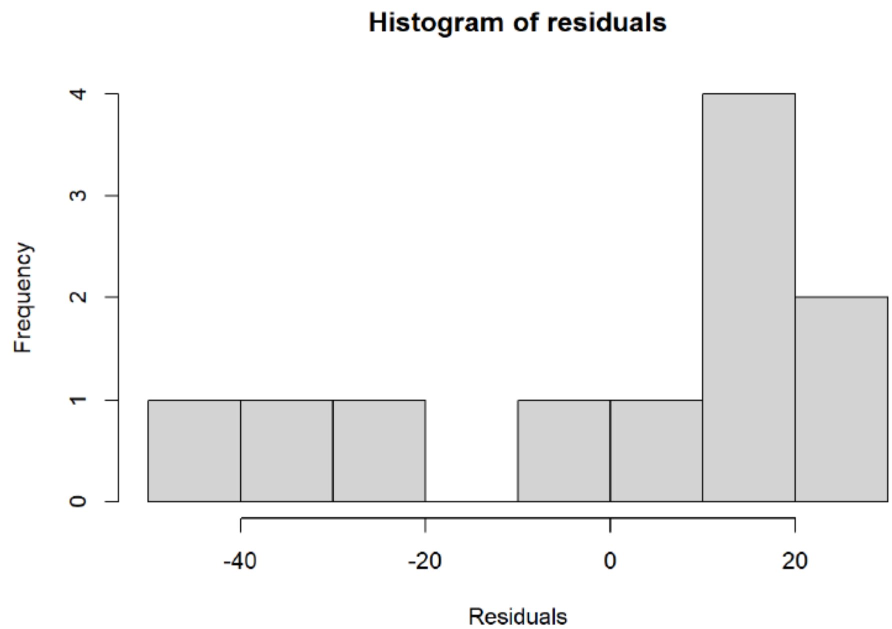
```
## Anova Table (Type III tests)
##
## Response: Depressed_Yes
##                                        Sum Sq Df F value  Pr(>F)
## (Intercept)                              8265  1  4.3461 0.12842
## Level_ofDepression1_3                    4806  1  2.5271 0.21013
## Year_TwoYearInterval                     8247  1  4.3365 0.12870
## TroubleSleeping_OrsleepingTooMuch         102  1  0.0539 0.83140
## FeelingTired_OrhavingLittleEnergy       38932  1 20.4720 0.02019 *
## Have_LittleInterestDoingThings           1643  1  0.8642 0.42114
## Feelingbad_aboutYourself                 2606  1  1.3705 0.32626
## Level_ofDepression1_3:Year_TwoYearInterval   4761  1  2.5036 0.21174
## Residuals                                5705  3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Anova Table (Type III tests)
##
## Response: Depressed_Yes
##                                        Sum Sq Df F value  Pr(>F)
## (Intercept)                            482225  1  9.7856 0.01665 *
## Level_ofDepression1_3                   13578  1  0.2755 0.61585
## Year_TwoYearInterval                   487140  1  9.8853 0.01629 *
## Level_ofDepression1_3:Year_TwoYearInterval  13310  1  0.2701 0.61930
## Residuals                              344955  7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for interaction term (level of depression:year) is non-significant ($p > 0.05$). So, there is no interaction between the variables of level of depression and year.

3.) The variance should be similar for all groups of the independent variable – homogeneity of variance

```
#testing assumption of homogeneity of variances
#null hypothesis: "there is no difference in variance between sample groups"
res1<-depression.model$residuals
#checking distribution of residuals to determine which homogeneity of variances test to conduct (Bartlett or Leve
ne)
hist(res1,main="Histogram of residuals",xlab="Residuals")
```




```r
#Checking normality among residuals through The Shapiro-Wilk test
#Null hypothesis: data is drawn from a normal distribution
shapiro.test(resid(aov(depression.model, data = finalProject_data)))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(aov(depression.model, data = finalProject_data))
## W = 0.88724, p-value = 0.1284
```

```r
#use bartlett test because data has been tested to be normally distributed but have slight inconsistencies
#in distribution of residuals not enough to say not normally distributed though
library(car)
```

```
## Warning: package 'car' was built under R version 4.1.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.1.3
```

```r
bartlett.test(Depressed_Yes ~ Level_ofDepression1_3, data = finalProject_data)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Depressed_Yes by Level_ofDepression1_3
## Bartlett's K-squared = 0.71526, df = 1, p-value = 0.3977
```

The p value for our Bartlett test is below 0.05 (0.3977<0.05). So, we fail to reject the null hypothesis, and conclude that each group of the level of depression variable have equal variances.

One-way ANCOVA Analysis

```
#Peforming One-way ANCOVA Test on our depression model
anova_test(data = finalProject_data, formula=Depressed_Yes~ Level_ofDepression1_3 + Year_TwoYearInterval+
             TroubleSleeping_OrsleepingTooMuch + FeelingTired_OrhavingLittleEnergy+
             Have_LittleInterestDoingThings + Feelingbad_aboutYourself+
             Level_ofDepression1_3:Year_TwoYearInterval,
           type = 3, detailed = TRUE) # type 3 SS should be used in ANCOVA
```

```
## Coefficient covariances computed by hccm()
```

```
## ANOVA Table (type III tests)
##
##                                       Effect       SSn     SSd DFn DFd      F
## 1                                (Intercept)  2207.973 5705.13   1   3  1.161
## 2                      Level_ofDepression1_3  4805.726 5705.13   1   3  2.527
## 3                       Year_TwoYearInterval  2236.137 5705.13   1   3  1.176
## 4          TroubleSleeping_OrsleepingTooMuch   102.441 5705.13   1   3  0.054
## 5          FeelingTired_OrhavingLittleEnergy 38931.724 5705.13   1   3 20.472
## 6             Have_LittleInterestDoingThings  1643.404 5705.13   1   3  0.864
## 7                   Feelingbad_aboutYourself  2606.288 5705.13   1   3  1.370
## 8 Level_ofDepression1_3:Year_TwoYearInterval  4761.111 5705.13   1   3  2.504
##       p p<.05   ges
## 1 0.360       0.279
## 2 0.210       0.457
## 3 0.358       0.282
## 4 0.831       0.018
## 5 0.020     * 0.872
## 6 0.421       0.224
## 7 0.326       0.314
## 8 0.212       0.455
```

Our ANCOVA results indicate there are significant differences in mean count of people who face depression ($p=0.02 < 0.05$) among people who have symptoms of feeling tired or having little energy throughout the day. This finding happens, while controlling the effects of the other covariate depression symptoms and adjusting the effect of people's depression levels.

Ultimately this suggests that the covariate depression symptom "feeling tired or having little energy throughout the day" is an important predictor of a person's depression level among the people collected in the NHANES sample population.

Additional Findings

Using the emmeans function test, we were able to discover the adjusted means for each group in the depression level variable. And surprisingly people who scaled their depression level a 1 out of 3 had a higher mean frequency count of facing the symptoms "feeling tired or having little energy throughout the day" than people who scaled their depression level a 2 out of 3.

Emmeans function test shown below:

```
#install.packages('emmeans')
library(emmeans)
```

```
## Warning: package 'emmeans' was built under R version 4.1.3
```

```
#getting estimated marginal means also known as  least-squares means
#for statistically significant covariate variable - FeelingTired_OrhavingLittleEnergy
adjustMeans <- emmeans_test(data = finalProject_data,
                            formula = Depressed_Yes ~ Level_ofDepression1_3,
                            covariate = FeelingTired_OrhavingLittleEnergy)
get_emmeans(adjustMeans)
```

```
## # A tibble: 2 x 8
##   FeelingTired_Or~ Level_ofDepress~ emmean    se    df conf.low conf.high method
##              <dbl> <fct>             <dbl> <dbl> <dbl>    <dbl>     <dbl> <chr>
## 1              365 1                  761.  28.6     8     695.      827. Emmea~
## 2              365 2                  678.  32.5     8     604.      753. Emmea~
```

<u>Post-Hoc Test</u>

Even though our categorical grouping variable “level of depression” was not statistically significant in our ANCOVA test of the model, I still ran a post hoc test on the formulated data. And the post-hoc test ultimately, further confirmed that there was no statistically significant differences in the depression mean frequency count among people with different depression levels.

```
#perform the post-hoc test with the Benjamini-Hochberg Hockberg method
emmeans_test(data = finalProject_data,
             formula = Depressed_Yes~ Level_ofDepression1_3,
             covariate = FeelingTired_OrhavingLittleEnergy,
             p.adjust.method = "hochberg")
```

```
## # A tibble: 1 x 9
##   term             .y.    group1 group2    df statistic     p p.adj p.adj.signif
## * <chr>            <chr>  <chr>  <chr>  <dbl>     <dbl> <dbl> <dbl> <chr>
## 1 FeelingTired_Or~ Depre~ 1      2          8      1.62 0.144 0.144 ns
```

Works Cited

Depression Data gathered from:
https://wwwn.cdc.gov/nchs/nhanes/default.aspx
(Continuous NHANES - Questionnaire Data Section of each year)

One-Way & Two-Way ANOVA Model examples:
https://dzchilds.github.io/stats-for-bio/one-way-anova-in-r.html

ANCOVA Analysis Introduction & Examples:
https://www.statology.org/ancova/

ANCOVA Assumptions:
https://r.qcbs.ca/workshop04/book-en/analysis-of-covariance-ancova.html#running-an-ancova

ANCOVA Analysis using R and Python:
https://www.reneshbedre.com/blog/ancova.html