Is There an Indicator For Depression: Prevalence of Depression in Different Groups

Terence Leung

March 26th, 2023

1 Introduction

Depression is a common and debilitating mental illness that affects millions of people worldwide. It can have a significant impact on an individual's quality of life, productivity, and overall well-being. Despite its prevalence, there is still much that is unknown about depression. The purpose of this paper is to examine whether there is a reliable indicator for depression across various groups, including race, age and other lifestyle factors such as obesity, alcohol abuse and drug abuse.

2 Methodology

The *health.dta* dataset contains 48784 observations and 155 variables. Each observation represents a patient record and each variable provides information about the patient, with personal details and indicator variables for a host of medical conditions. In order to determine the relationship between demographic with depression, I will first create a new variable that gives a rough age rather than an age range so that the regression would make more sense. Because the depression indicator is a binary variable, with 1 indicating the patient suffers from depression, the results of the regression are simply to interpret. Note that in all the following regressions, the binary depression indicator is always the dependent variable and we will regress the depression indicator on various lifestyle and demographic indicators. For the age group regression, the coefficient means that for each year increase in age, we expect the that much more of the population corresponding to that age to experience depression. The constant term relates to the percent of people experiencing depression at age equals to 0. The regression for race is even easier to interpret since both variables are binary - the coefficient shows how much more of the black population is affected than the white population and the constant term shows how much of the white population is affected. This allows us to answer the question of whether demographic group is a strong indicator for depression.

For the second part of our analysis, we will use the same approach and run regressions for lifestyle factors such as obesity, alcohol abuse and drug abuse. From this, we will know how much of the obese population is depressed versus the non-obese population, and the same for the other two factors. It is worth noting that these comparisons do not reflect any causal relationships, as it is just as possible that depression causes certain lifestyle choices as certain lifestyle choices cause depression.

3 Results

We obtained the following result from the age group regression.

```
      Source |       SS           df       MS      Number of obs   =    47,450
-------------+----------------------------------   F(1, 47448)     =    330.55
       Model |   19.206941         1   19.206941   Prob > F        =    0.0000
    Residual |  2757.01917    47,448   .05810612   R-squared       =    0.0069
-------------+----------------------------------   Adj R-squared   =    0.0069
       Total |  2776.22611    47,449  .058509686   Root MSE        =    .24105

------------------------------------------------------------------------------
tm1_depres~r | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
   age_group |   .0013223   .0000727    18.18   0.000     .0011797    .0014648
       _cons |  -.0061101   .0039275    -1.56   0.120     -.013808    .0015878
------------------------------------------------------------------------------
```

The results are highly significant ($p < 0.001$). Since the age groups were separated to every 10 years, the coefficient of 0.0013223 means that we can expect each age group to have 1.32% more of its population be affected by depression. The constant coefficient -0.00611 means that 2.03% of the first age group of 20 is affected by depression, 3.35% of the second age group of 30 is affected etc. The R-squared of this regression is 0.0069, meaning only 0.69% of the variance in population affected by depression is explained by this model.

Performing the race regression yields.

```
      Source |       SS           df       MS      Number of obs   =    47,450
-------------+----------------------------------   F(1, 47448)     =     73.74
       Model |  4.30803346         1  4.30803346   Prob > F        =    0.0000
    Residual |  2771.91808    47,448  .058420125   R-squared       =    0.0016
-------------+----------------------------------   Adj R-squared   =    0.0015
       Total |  2776.22611    47,449  .058509686   Root MSE        =     .2417

------------------------------------------------------------------------------
tm1_depress~r | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
--------------+---------------------------------------------------------------
```

```
tm1_dem_black |   .0299894   .0034923     8.59   0.000     .0231445    .0368343
        _cons |   .0589858   .0011788    50.04   0.000     .0566754    .0612962
------------------------------------------------------------------------------
```

The results are highly significant ($p < 0.001$). From the coefficients, we can conclude that 5.90% of the white population experiences depression while 8.90% of the black population experiences depression. However, the R-squared value of 0.0016 would indicate that only 0.16% of the variance is explained by this model.

Diagram 1 below summarizes the results - there is a general trend for both races that depression becomes more prevalent as they get older. Further, it is also generally true that blacks suffer more from depression than whites at nearly all age groups. Blacks in the age group of 65-74 suffer the most from depression, affecting about 14% of all black people between the ages of 65-75

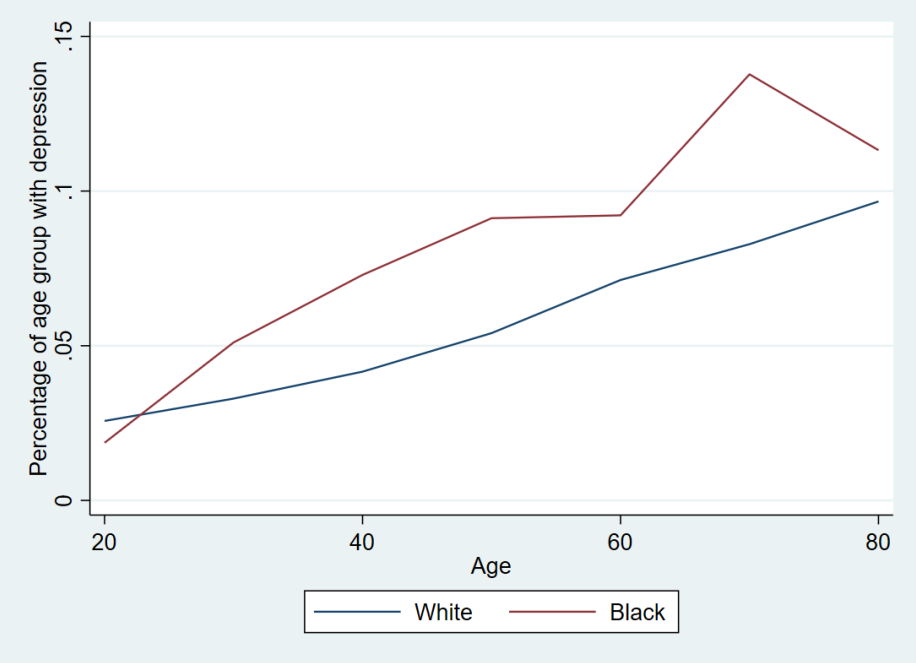

Diagram 1 - Twoway line graph comparing prevalence of depression different age groups of white people and black people

The results for the lifestyle factors were all highly significant ($p < 0.001$). For the sake of brevity, the results can be summarized as follows: 16.8% of the obese population has depression vs 5.13% of those who are not obese; 26.4% of those with signs of alcohol abuse vs 6.02% of those without; 36.5% of those with signs of drug abuse vs 6.02% of those without. When running the regressions with these three factors, the R-squared values were 0.0193, 0.0063 and 0.0099 respectively, meaning out of the three factors, the model that obesity is an indicator for depression is strongest. Diagram 2 below summarizes the findings.

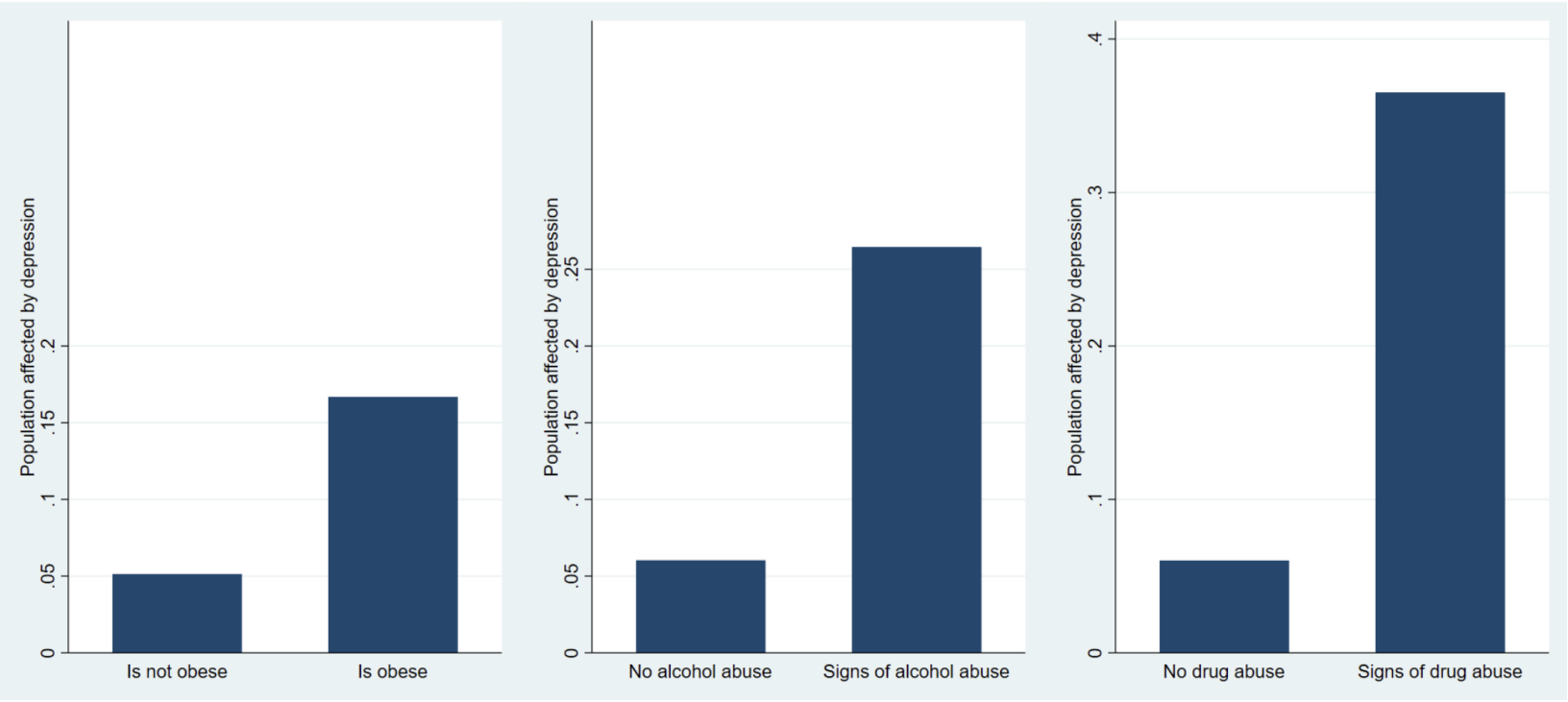

Diagram 2 - Prevalence of depression among people who are obese, has signs of alcohol abuse or drug abuse, and those who do not

By comparing R-squared values between the various regression models, we can conclude that the obesity model is the strongest while the race model is the weakest, while the other three factors of age, alcohol and drug abuse hold about the same strength.

4 Conclusion

From our results, we can determine that the examined factors are all positively correlated with depression. Since there are so many factors that can contribute to whether a person experiences depression, we cannot say absolutely how strong these factors are as an indicator of depression. However, we know that relative to each other, the obesity model is the strongest while the race model is the

weakest. A big flaw of this experiment is that the R-sqaured values of examined factors were all very low, and it is hard to determine how relevant they are given that there are so many factors to consider.

Through this exercise, I have learned how to effectively explore, analyze and visualize a research question using Stata.

This paper contributes to the literature as a starting point to examining the indicators of depression. Future research could focus on identifying causal relationships between lifestyle factors and depression. Or alternatively, explore the impact of other demographic factors such as income, education, and social support on depression.