Analysis of People Diagnosed with Depression

Samahriti Mukherjee[1] , Scott Brinley [2] , Nina Bryan[2]

[1]Indian Statistical Institute, [2]University of Michigan

Importance of Our Research

- Depression often occurs with other illnesses and medical conditions:
  - 25% of cancer patients experience depression
  - Depression is the second most common mental health condition among patients living with HIV
- Over 20% of Americans with anxiety or mood disorders have an alcohol or other substance use disorder
- Students who smoked were 2-3x more likely to have clinical depression than those who have never smoked
- Women are twice as likely as men to have had depression
- Depression contributes to an estimated $100 billion annual cost for US employers, including $44 billion a year in loss of productivity alone.
- Depression impacts all aspects of life which include relationships with family and community, poor productivity and concentration, and disrupted sleep and eating habits.
- Statistical studies on depression help us determine the prevalence of depression in different genders, regions, demographic groups etc.
- This data helps us identify high risk groups and develop prevention strategies.

Description of the Data

MGI Dataset :

- Training Set: 46,154 Observations,
- Testing Set: 21,066 Observations
- Classification/Labels: 2 (Non-Depression (0) , Depression (1))
- # of Features: 83

Objective & Methodology

To predict and do a Causal Inference on the Depression diagnosis of patients in the MGI dataset based on a combination of different feature variables including recorded comorbidities and demographic characteristics.

Methodology:

- Classification:

We have used Logistic, Naïve Bayes, LDA, QDA, SVM (Linear, Radial), Decision Tree, Random Forest, Boosting, Lasso, Ridge and Ensemble Learning to classify people diagnosed with Depression and AUC to analyze performance of each algorithm.

- Causal Inference:

We have used IPW, AIPW and Direct Estimation for calculating ATE and to analyze Heterogeneity in Treatment Effect.

Mathematical Formula

- **Classification:**

We predict the probability of assigning a person to the class 1 or 0 as

$$\hat{P}(\text{Depression}_i = 1 | \vec{X}_i = \vec{x_i}) = \hat{f}(\vec{x_i})$$

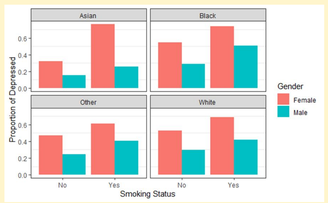
With feature vector $\vec{x} = \langle x_1, x_2, \dots, x_n \rangle$.

- **Causal Inference:**

We calculate the absolute standardized mean difference (ASMD) between treated and untreated individuals as $\frac{|\overline{Z_1} - \overline{Z_0}|}{\sqrt{s_1^2 + s_0^2}}$, where $\overline{Z_1}$, $\overline{Z_0}$ are sample averages and s_1, s_0 are standard deviations of Z_i (covariate) for the treated and untreated individuals. We calculate the same quantity for their weighted counterparts $\frac{Z_i W_i}{\hat{e}(X_i)}$ *and* $\frac{Z_i(1-W_i)}{(1-\hat{e}(X_i))}$.

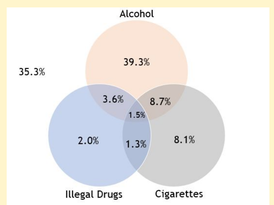
Visualization

- Depression Diagnosis w.r.t. Smoking Status, Gender & Races:



Based on our dataset, we can conclude that smokers are more frequently diagnosed with depression than non smokers in both males and females. Similar Case also arise by plotting the same for various races.

- Intersectionality of Risky Behaviors

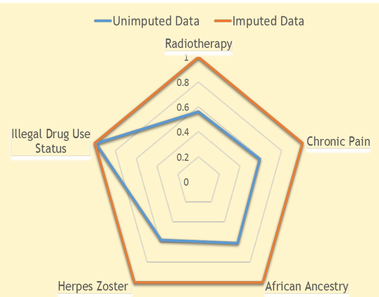


Of the patients diagnosed with depression, a large portion participate in risky behaviors, such as smoking cigarettes, drinking alcohol, and ingesting illegal drugs. This plot displays the importance of exploring trends in subgroups which may be become insightful when interpreting our results for each model. Especially, when substance abuse is highly correlated with depression in the US population.

Missing Value Imputation

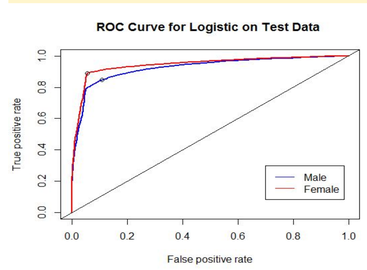
Merging multiple datasets led to large amounts of missing data and Imputation allows us to **keep large sample size.**

Used **MICE technique** to impute missing predictor and outcome variable.



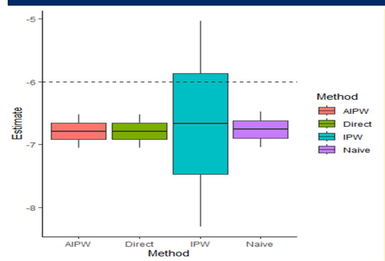
Prediction of Depression Diagnosis

As Logistic, LDA and Linear SVM are the best models in terms of AUC value, we can say that there is a linear relationship between being diagnosed with depression and other variables.

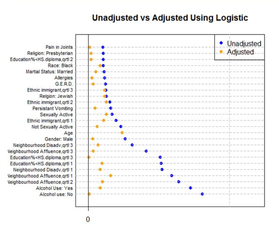


We see that Logistic Regression gives the best AUC value for both male and female. So we present a comparison of ROC curves for both male and female.

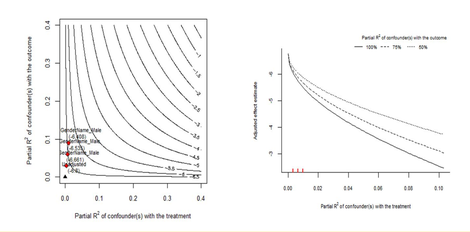
Causal Inference



The spread for IPW is the most as logistic is very much biased. Random Forest is also biased. But Lasso, XgBoost and Ridge are very close to the true value (-6), with Ridge giving the best result.



Weights are able to reduce the heterogeneity for both IPW and AIPW.



We fit a linear model to analyze how many times stronger the confounder is related to the treatment and the outcome in comparison to Male (Observed Benchmark Covariate) in explaining treatment outcome variation.

Future Research

- The contrast in feature importance between subgroups (Male and Female) needs to be further investigated and could ultimately impact treatment and intervention
- The notable interactions in the logistic model are heavily aligned with depression research in US population. Research on cigarette marketing or cultural habits and norms that influence partaking in risky behaviors (peer pressure)
- The research between women and chemotherapy, especially since studies have shown that women are more likely to be left by their partner when diagnosed with cancer
- Missing variable (irregular cycle) and impact on women prediction models
- The bias in alcohol drinkers, is it observational bias?

Acknowledgement: Dr. Rahul Ladhania, Ritoban Kundu, Dr. Bhramar Mukherjee

References

- Depression statistics. Depression and Bipolar Support Alliance. (2019, July 12). https://www.dbsalliance.org/education/depression/statistics/
- Goodwin, R. D., Dierker, L.C., Wu, M., Galea, S., Hoven, C.W., & Weinberger, A.H. (2022). Trends in US Depression Prevalence From 2015-2020: The Widening Treatment Gap. American journal of preventive medicine, 63(5):726-733. https://doi.org/10.1016/j.ampere.2022.05.014
- Class Slides