Depression on Drug Use

Vanessa Okosun

Prof. Arnab Nayak

**Fulfillment Statement**: A capstone project for partial fulfillment of the requirements

of Master of Science in the Business Analytics program offered by the Stetson-Hatcher School of

Business at Mercer University.

Depression and Drug Use

*Introduction*

According to the National Institute of Alcohol Abuse and Alcoholism (NIAAA), about 4% of Americans met the requirements of being categorized as having a drug dependency or drug problem. That means that about 23 million people in the United States have struggled with drugs at one point. I don’t know about you, but those statistics are pretty shocking. It is essential to dive into the many variables that may take into account why so many people, not only in the US but in the world, are dealing with these drug issues. With a public health background, I know prevention is always better than treatment, which influenced me to take on this subject.

My focus of this study is to explore different variables or factors that can influence whether a person is a drug user. I choose depression as the primary independent variable because I know that there is a correlation between them. Depression can affect many things, such as how you feel, social interactions, or how you interact with different daily activities. According to the National Survey on Drug Use and Health (NSDUH), about 21 million people in the United States have at least one episode of depression. As you can see, many people in the united states are equally dealing with depression and drug dependency. Drugs can cause chemical changes to the brain that can initiate depression. Depression can be passed down via genetics or entirely unrelated to genetics but influenced by life activities. Depression can potentially lead people to these different substances, but drug use can also lead to depression. The relationship between these two variables is bi-directional.

My goal is that diving into this topic with analytics would help me explore different internal and external factors, including depression, that can influence the prevalence of being a drug user. Taking the data and using analytical models can help solve this predictive question. When doing this, we can help figure out the different early preventive measures that can help decrease both the number of people dealing with depression, diagnosed and undiagnosed and the people dealing with drug dependency or use.

The drugs that I am focusing on in my study are marijuana, cocaine, methamphetamine, and heroin. The other external factors I could obtain in my data were gender, race, marital status, income, poverty ratio, age, and education. I will be discussing the literature review that I did to increase my knowledge of this topic, the data I obtained to pursue my goal, the cleanup process, my data mining process, and my conclusion and recommendations.

Once completing my model, I was able to see that depression and drug use were correlated along with other variables such as gender and martial status.

*Literature Review*

Before diving into the problem, improving my current knowledge of the topic is essential. There are a few articles that helped me scope out my process. I found a study that compared independent depression to substance-induced depression. The substances included in this research were marijuana, cocaine, and Opioids. The research question was, do depressive symptoms often coexist with substance use disorders? The advantage that I have in my dataset is that I was able to obtain a questionnaire that asks them about their drug history. Although my goal was not to pinpoint the type of depression, I could see the methodology behind the study. Similar external variables that I used were also used in the study. This study is where I got to use a multivariate logistic regression. This study helped me determine which variables were significant predictors of being a drug user.

Logistic regression was also used in another study that explored depression and substance abuse in middle school-age kids. It is a substantial public health issue that children are using these addictive substances at such a young age, and it is reported that the average age at that most kids start using drugs is around 14. The study dived into how depression and using these substances can affect grades in school. Similar to what I did, they created new variables to classify the students as drug users and depressed using the scoring system of the two different questionnaires.

One study that I did some research on allowed me to see how to incorporate a simultaneous model. In the study on the relationship of smoking and depression in Indonesia, they were able to see that in both men and women both of those variables occur simultaneously. Depression can be seen as both a result and cause of smoking or in my case, drug use. Due to this, there is bias in the model causing the dependent variable’s error terms are correlated with the independent variables. Simultaneous model helps correct that bias by using two-stage least square. Depression and smoking was broken up into two equations. Depression and smoking being the dependent in their own equations. In each of the equations, the other variable is used with exogenous variables that are used for prediction.

*Data, data sources, and data characteristics* (see ***Data Analysis Appendix****)*

I obtained the dataset from the National Health and Nutrition Survey (NHNS), which I found through the CDC page. The NHNS is a program of research that evaluates children's and adults' nutritional status and overall health in the United States. Its responsibility is to produce vital health statistics for the country. The sample of the survey that I obtained was 2017-2018 data. I carefully had to consider which variables would be essential to have for my study. I obtained the drug use, mental health-depression screener, and demographics data. These datasets are all connected by a sequence number. The sequence number is an "identifying" variable for everyone screened. I used Microsoft Access to combine the dataset by setting the sequence number as the primary key.

There were many variables included in the dataset after being combined. The dataset initially consisted of 4572 observations and 105 variables (see appendix a). My first main concern was clean up the demographics because there were a lot of unwanted variables that would not help achieve the goal of my model. The only variables that I ended up keeping from that sector are sequence number(SEQN), gender(Gender), marital status(MARTL), education(EDUC), income, poverty ratio, and age in a year. My input variable was if they were depressed or not, and all the demographic variables that I kept. The output variable will be if they are a drug user or not.

The questionnaire for drug use included questions such as, "Have little interest in doing things," "Have little interest in doing things?", and "how often have you been bothered by the following problems:] feeling down, depressed, or hopeless?". The scale goes from 1 being severely days/somewhat difficult and 3 being nearly every day/extremely difficult. The questionnaire for the mental health-depression screening includes questions such as, "Ever used marijuana or hashish," "Ever used cocaine/heroin/methamphetamine," "Ever use a needle to inject illegal drug" and "Have ever been in a rehabilitation program".

*Data Analysis Preparation*

Due to the dataset including so many variables, I had to find a way to decrease the variables I wanted to focus my model on. I decided to create a drug user and depression index. I created a scale for depression. Normal is from 0-4, minimal is 5-9, mild is 10-14, moderate is 15-19, and moderately severe depression is a score of over 19. In this model, I considered those who have mild, moderate, and moderate depression to be depressed. For drug use, there were many variables to consider. However, only six were used for the drug user classification (if marijuana was used every month for a year, if they used marijuana more than twice a month, if cocaine was used more than two times, if they used heroin, if they used methamphetamine more than twice, if they have ever used a needle to inject illegal drug and if they ever been in rehab program). By decreasing the data dimension, I could decrease from 105 variables to 52. That count also included the index and drug price variables I created.

All of the variables in the dataset were dummy variables. Once I loaded the dataset into R, I first decreased the dummy variables into two, 0 and 1. This applied to race (black, non-Hispanic -1, other-0), education level( 0 – high school equivalent less than; 1 - college and above), marital status (0 - not married; 1 – married), and income level (0- under $20,000, 1 - over $20,000). I also added variables for drug prices for marijuana, heroin, methamphetamine, cocaine, and cocaine. If they used the drug, the price was added. If they did not, a 0 was inputted.

The data mining task is to predict whether a person will become a drug user based on if they have depression along with other factors mentioned before. Depression, gender, race, education, marital status, and poverty ratio will be the independent variables, and being a drug user is the dependent variable. The data was split between validation and training. 70% would be training data, and 30% would be validation data. To achieve my goal of prediction, I used logistic regression and improved it by doing a simultaneous equation model. A logistic regression is appropriate because it relates predictors to an outcome. It helps find the odds of an event happening. The simultaneous equation model helps improve simultaneity. Typically, x is independent, and y is dependent. In this case, x and y are dependent on each other. The relationship between depression and drug use is bi-directional. The simultaneous equation model breaks down my equation into two separate regressions using predictors precise to depression and being a drug user. Although I am using this method, the only weakness I will encounter is whether this is a good predictive model.

*Empirical Results (see* ***Results Appendix****)*

The logistic regression was the first test that I applied. I created two models to help find the output of being a drug user using the training data that consisted of 70% of the data. The first model was the influence of being depressed, gender, marital status, race, education, and the poverty ratio on being a drug user. I got the p-value of each of the independent variables. The p-value tells you if the variable is statistically significant in the change of the dependent variable. It is used in hypothesis testing. The null hypothesis is that the variable does affect the sample population. The alternative hypothesis means that there is an effect on the population. A small p-value means that we will reject the null hypothesis meaning that it is statically significant. The only significant variables in the first model were the depressed, gender, marital status, and race. I then used the significant variables and created model 2. I was able to compute each variable's importance. According to the results, marital status is the most important variable in the prediction of being a drug user. Gender is the second, depressed is the third, and race is the fourth.

I ran a model comparison and looked at the AIC weight to see which model had a higher predictor power. The AIC weight for model 2 was 82% compared to model 1, with only 18% of predictive power. I could convert the coefficients from the logistic regression to an odd ratio and then a percentage. Those who are depressed are 49% more likely to be a drug users. Females are 48% less likely to be drug users. African Americans are 29% more likely to be drug users, and those who are married are 57% less likely to be drug users. My McFadden's R2 value is .058. The closer this value is to 1, the higher the predictive power. Unfortunately for my model, it has a low predictive power, but I can still see the correlation between being depressed and being a drug user.

I created two new samples of individuals to make predictions with my model. Individual one is an African American, unmarried female with an education of high school or below who is depressed. Individual two is an African American, unmarried female with an education of high school or below who is not depressed. A woman who is depressed has a 37% chance of being a drug user. A woman who is not depressed has a 24% chance of being a drug user. To test my model performance, I calculated the AUC. It is the area under the curve that tells you how well my model does at classification. My AUC is .675. The closer this value is to 1, the better the model is at classification or prediction. I also created a lift chart in R. The lift chart is for a sample of 1400 people. The model's predictive performance in terms of lift is better than the baseline model since its lift curve is higher than that of the baseline model.

In my results, I see a correlation between being depressed and being a drug user. My model is simultaneous. Changes are happening to both the depressed and drug user variable simultaneously. Typically, x is the independent variable, and y is the dependent variable. In this specific case, depression and drug use depend on each other, meaning they have a bi-directional relationship. Having this in your model can lead to errors in my estimates. Depressed and drug users are endogenous variables. The exogenous variables are the indicator variables for the demographics of the individual. The variables specific to being a drug user are the prices of the drug. If the individual is classified for that specific drug, the price per gram was inserted for marijuana, cocaine, meth, and heroin. The variables that are specific to being depressed are gender, marital status, and income. All of the variables for both depressed and drug users are significant. The 2SLS was compared to the 3SLS. Hausman's specification test tells me to reject the 3SLS because the 2SLS is better.

*Conclusions and Recommendations*

Depression and drug use are highly correlated. It was also interesting to see how other external factors come into play, like marital status and gender. Even though my model did not yield a high predictive power, it was good to see this correlation. This would help with prevention and early intervention methods. Compared to other publications that had similar goals, we were able to get the same results. Suppose I got more data that are more specific to demographics. For future purposes, I would love to see how the age of the first diagnosis of depression and drug use plays a part in this study. More research in this area will help boost protective measures and eliminate those risk factors for drug use. I also believe that mental health checkups should be included in yearly physicals. This will allow health professionals to catch it early to avoid severe depression or drug use due to depression.

References

Depression. (n.d.). National Institute of Mental Health (NIMH). https://www.nimh.nih.gov/health/topics/depression

Major Depression. (n.d.). National Institute of Mental Health (NIMH). https://www.nimh.nih.gov/health/statistics/major-depression

Villarroel, M., & Terlizzi, E. P. (2020). Symptoms of Depression Among Adults: United States, 2019. NCHS Data Brief, 379, 1–8. https://pubmed.ncbi.nlm.nih.gov/33054920/

Bustamante, J. (2023, January 1). Drug Abuse Statistics. NCDAS. https://drugabusestatistics.org/

NHANES - About the National Health and Nutrition Examination Survey. (n.d.). https://www.cdc.gov/nchs/nhanes/about\_nhanes.htm

Income inequality and depression: a systematic review and meta-analysis of the association and a scoping review of mechanisms. (n.d.). https://onlinelibrary.wiley.com/doi/full/10.1002/wps.20492

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl  Jr., K. C. (2018). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*.

Spector, N. (2018, January 22). Why don’t Americans get regular mental health checkups? It’s complicated. NBC News. <https://www.nbcnews.com/better/health/why-aren-t-mental-health-screenings-part-our-annual-physicals-ncna839226>

Kelder, S. H., Murray, N., Orpinas, P., Prokhorov, A. V., McReynolds, L. S., Zhang, Q., & Roberts, R. (2000). Depression and substance use in minority middle-school students. American Journal of Public Health, 91(5), 761–766. <https://doi.org/10.2105/ajph.91.5.761>

Henningsen, A., & Hamann, J. D. (2007). **systemfit**: A Package for Estimating Systems of Simultaneous Equations in*R*. *Journal of Statistical Software*, *23*(4). https://doi.org/10.18637/jss.v023.i04

Dakwar, E., Nunes, E. V., Bisaga, A., Carpenter, K., Mariani, J. J., Sullivan, M. A., Raby, W. N., & Levin, F. R. (2011). A Comparison of Independent Depression and Substance-Induced Depression in Cannabis-, Cocaine-, and Opioid-Dependent Treatment Seekers. American Journal on Addictions, 20(5), 441–446. <https://doi.org/10.1111/j.1521-0391.2011.00148.x>

Grant, B. F. (1994). Comorbidity between DSM-IV drug use disorders and major depression: Results of a national survey of adults. Journal of Substance Abuse, 7(4), 481–497. <https://doi.org/10.1016/0899-3289(95)90017-9>

McCarty, C. A., Wymbs, B. T., Mason, W. A., King, K. M., McCauley, E., Baer, J., & Stoep, A. V. (2013). Early Adolescent Growth in Depression and Conduct Problem Symptoms as Predictors of Later Substance Use Impairment. *Journal of Abnormal Child Psychology*, *41*(7), 1041–1051. <https://doi.org/10.1007/s10802-013-9752-x>

Kelder, S. H., Murray, N., Orpinas, P., Prokhorov, A. V., McReynolds, L. S., Zhang, Q., & Roberts, R. (2000b). Depression and substance use in minority middle-school students. *American Journal of Public Health*, *91*(5), 761–766. <https://doi.org/10.2105/ajph.91.5.761>

Liew, H., & Gardner, S. (2016). The interrelationship between smoking and depression in Indonesia. *Health Policy and Technology*, *5*(1), 26–31. <https://doi.org/10.1016/j.hlpt.2015.10.003>

Appendix

*Data Analysis Appendix*

Table 1.1

|  |  |  |
| --- | --- | --- |
| Variable Name | Description | Data Type |
| SEQN | Respondent sequence number | Integer |
| Gender | Gender of the participant | Binary |
| AGEYR | Age in years at screening | Binary |
| RIDRETH1 | Race/Hispanic origin | Binary |
| EDUC | Education Level | Binary |
| MARTL | Marital Status | Binary |
| POVRAT | A ratio of family income to poverty guidelines. | Integer |
| INCOME | Annual household income | Binary |
| DPINTEREST | Have little interest in doing things | integer |
| DPFEEL | how often have you been bothered by the following problems:] feeling down, depressed, or hopeless? | integer |
| DPSLEEP | Trouble sleeping or sleeping too much | integer |
| DPENERGY | Feeling tired or having little energy | integer |
| DPAPPETITE | Poor appetite or overeating | integer |
| DPSELF | Feeling bad about yourself | integer |
| DPCONCEN | Trouble concentrating on things | integer |
| DPSPEAK | Moving or speaking slowly or too fast | integer |
| DPTHOUGHT | Thought you would be better off dead | integer |
| DPDifficulty | Difficulty these problems have caused | integer |
| DUQ200 | Ever used marijuana or hashish | integer |
| DUQ210 | Age when first tried marijuana | integer |
| DUQ211 | Used marijuana every month for a year? | integer |
| DUQ213 | Age started regularly using marijuana | integer |
| DUQ215Q | Time since last used marijuana regularly | integer |
| DUQ210U | Time since used marijuana regularly/unit | integer |
| DUQ217 | How often would you use marijuana? | integer |
| DUQ219 | How many joints or pipes smoke in a day? | integer |
| DUQ220Q | Last time used marijuana or hashish | integer |
| DUQ220U | Last time used marijuana or hashish/unit | integer |
| DUQ230 | # days used marijuana or hashish/month | integer |
| DUQ240 | Ever used cocaine/heroin/methamphetamine | integer |
| DUQ250 | Ever use any form of cocaine | integer |
| DUQ260 | Age first used cocaine | integer |
| DUQ270Q | Last time you used cocaine, in any form | integer |
| DUQ270U | Last time you used cocaine/unit | integer |
| DUQ272 | # of time you used cocaine | integer |
| DUQ280 | # of days used cocaine/month | integer |
| DUQ290 | Ever used heroin | integer |
| DUQ300 | Age first used heroin | integer |
| DUQ310Q | Last time used heroin | integer |
| DUQ310U | Last time used heroin/unit | integer |
| DUQ320 | # of days used heroin/month | integer |
| DUQ330 | Ever used methamphetamine | integer |
| DUQ340 | Age first used methamphetamine | integer |
| DUQ350Q | Last time used methamphetamine | integer |
| DUQ350U | Last time used methamphetamine/unit | integer |
| DUQ352 | # times used methamphetamine | integer |
| DUQ360 | # days used methamphetamine/month | integer |
| DUQ370 | Ever use a needle to inject illegal drug | integer |
| DUQ380A | Drugs injected - Cocaine | integer |
| DUQ380B | Drugs injected - Heroin | integer |
| DUQ380C | Drugs injected - Methamphetamine | integer |
| DUD380F | Drugs injected - Steroids or other drugs | integer |
| DUQ390 | Age first injected drugs | integer |
| DUQ400Q | Last time injected drugs | integer |
| DUQ400U | Last time injected drugs/unit | integer |
| DUQ410 | # times injected drugs/lifetime | integer |
| DUQ420 | How often did you inject drugs | integer |
| DUQ430 | Ever been in rehabilitation program | integer |

Chart 1.1 – Age distribution

*Chart, histogram

Description automatically generated*

Chart 1.2 – Education Counts

Chart, bar chart

Description automatically generated

Chart 1.3 – Gender Counts ( 1 – male, 2-female)

*A picture containing chart

Description automatically generated*

Chart 1.4 – Race Counts (Black Non-Hispanic – 1)

*Chart, bar chart

Description automatically generated*

Chart 1.5 – Married Counts

A picture containing chart

Description automatically generated

Chart 1.6 – Depressed Counts

Chart, bar chart

Description automatically generated

Chart 1.7 – Drug User Counts

Chart, bar chart

Description automatically generated

Table 1.2 – Depression Scale

|  |  |
| --- | --- |
| Score | Meaning |
| 1 | Several days/somewhat difficult |
| 2 | More than half the days/very difficult |
| 3 | Nearly every day/extremely difficult |
| 7 | Refused |
| 9 | Don't know |
| . | Missing |

*Results Appendix*

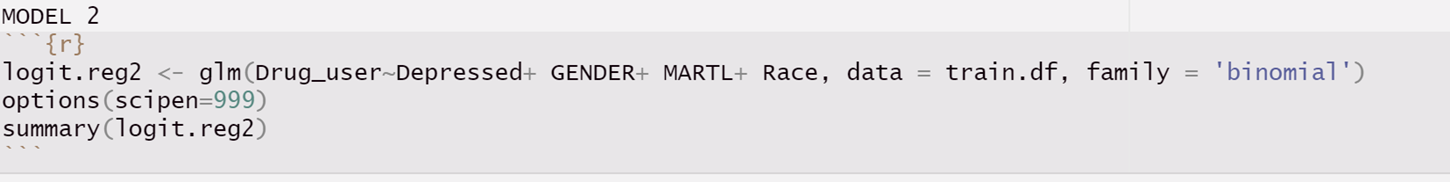
3.1 – logistic regression, model 1

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

3.2 – logistic regression, model 2

Table

Description automatically generated

3.3 – model comparison

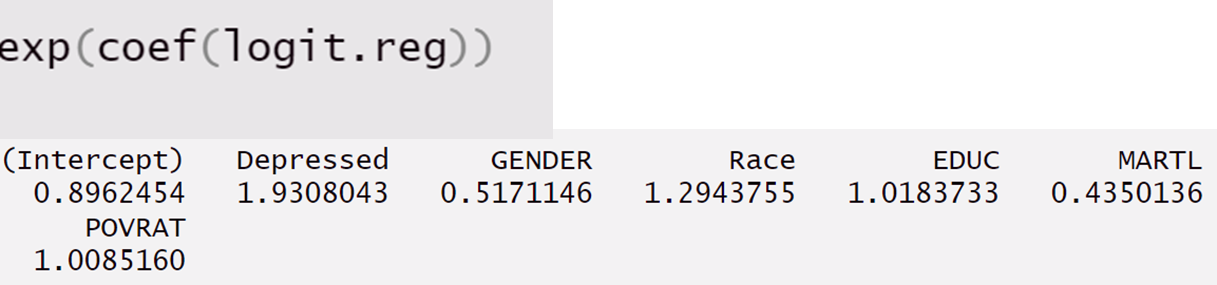
Text

Description automatically generated with medium confidence

3.4 – model validation 1

Graphical user interface, text, application, email

Description automatically generated



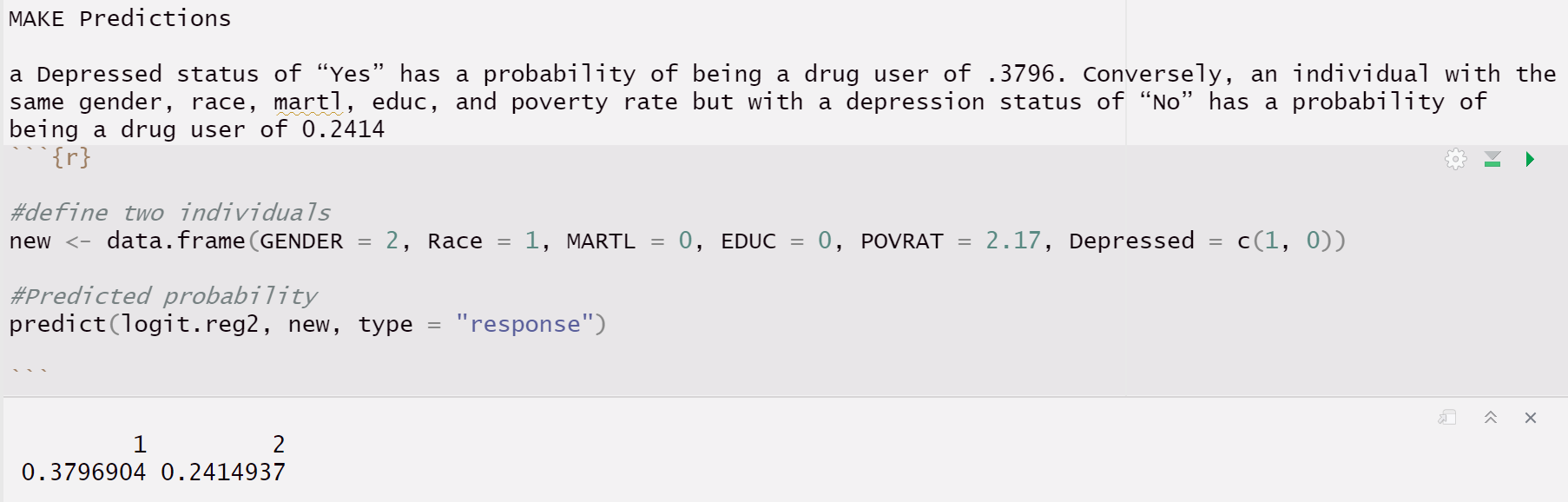
3.5 – variable importance

Graphical user interface, application, table

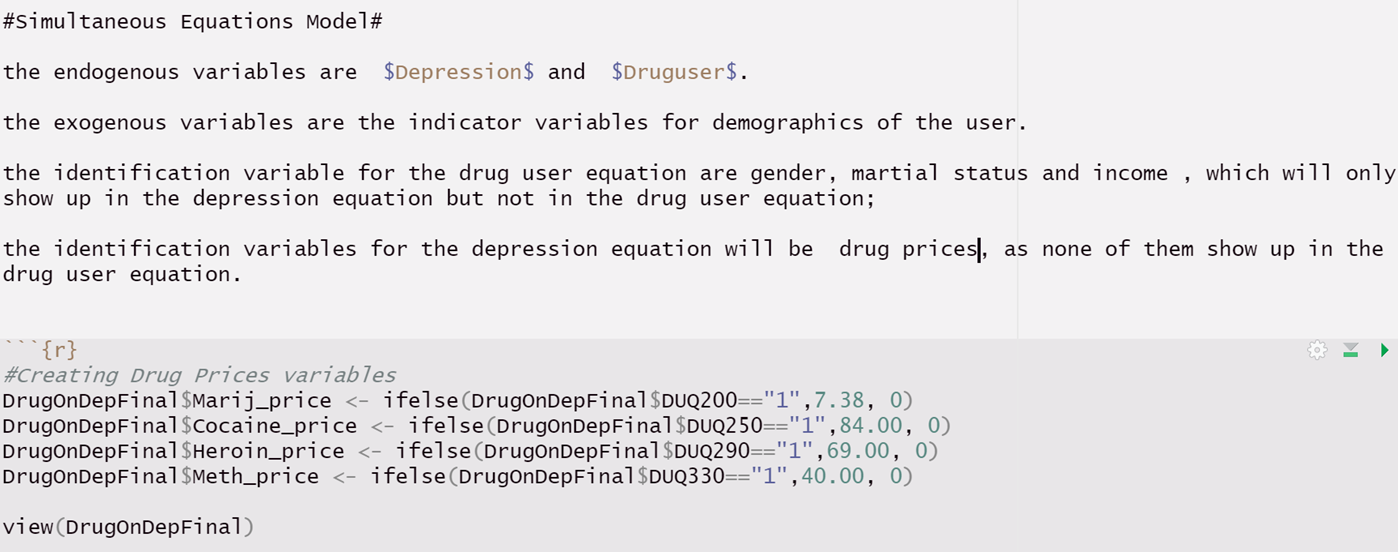
Description automatically generated



3.6 – predictions



3.7 - Simultaneous Equations Model



Graphical user interface, text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

3.8 – 3sls

Text, letter

Description automatically generated

Text, table

Description automatically generated

Text

Description automatically generated

3.9 – Hausman test

Graphical user interface, text, application, email

Description automatically generated