Predicting mental health outcomes

Julia Lee

1 Introduction

More than 20% of Americans have a brain or mental health disorder[1]. The direct causes of mental illness are unknown. The earlier individuals are diagnosed with a mental disorder, the sooner they can get help before it becomes more severe[2]. Therefore, it is important to know what factors would increase someone's risk of mental illness. Studying risk factors of mental illness will help governments, communities and medical professionals to be able to determine where to best distribute resources for high risk communities. This paper analyzes how past and current sociological and physical factors influence the current trajectories of mental health of individuals.

This analysis is based on data from The Midlife in the United States (MIDUS) study. The MIDUS study is a longitudinal study that investigates the physical health, psychological well-being, and social responsibility of individuals in middle life. The data was collected in three waves of repeated measures on individuals under MIDUS. The first wave was funded by the MacArthur Foundation Research Network on Successful Midlife Development. The second and third wave of research was funded by grants from the National Institute on Aging (P01AG020166, R37AG027343) [3]. All participants were non-institutionalized, English-speaking adults in the coterminous United States, aged 25 to 74. We are going to analyze a subset of the data, focusing on the first wave (1995-1996) and the second wave (2005-2006). The data set has 3439 observations and 56 variables. The variable names that start with A are from the first wave and B are from the second wave.

One of the variables that MIDUS study measures is the mental health of the individuals. They have a binary variable that measures whether the individual was treated for /diagnosed with / or experienced anxiety, depression or other mood disorders in the past 12 months.

In this paper we will use various methods to analyze this data to investigate the risk factors of mental illness over time. We will compare and contrast these methods to explore which method or methods are best for analyzing the data. The methods that will be used to explore this problem are decision trees, logistic regression, smoothing splines and the Lasso Method for variable selection.

2 Exploratory Data Analysis

2.0.1 Data Transformation

Initially we filtered out missing data, leaving a new data set with 3104 observations. We created a variable for age by transforming birth year into a number by subtracting the birth year from 2006 and from 1996. In order to standardize our data to include variables that measure change over time, we created some new variables. We created a variable that tells us about whether a person's mental health status changed from the first to the second wave. We also created binary variables that tell us about whether a person's highest level of education and whether their marital status changed from the first to the second wave. We also wanted to include the interaction between factors–for example we created a variable that is the interaction both neurological measurement factors this creates 3 factor levels (1, 2, 4).

In order to account for the relationship between past and current mental health, we broke the data into 2 subsets: one data set where the individuals did not previously report mental illness and another data set where the individuals did previously report mental illness.

[1]"Mental Disorders in America." Mental Illness Statistics. N.p., n.d. Web. 07 May 2013.

[2]Tarugu, Vikram. "The Real World: Recognizing Mental Illness in Young Adults." NBC 5 Dallas-Fort Worth, NBC 5 Dallas-Fort Worth, 12 July 2009

[3]Http://Midus.wisc.edu/Findings/Pdfs/1387.Pdf.

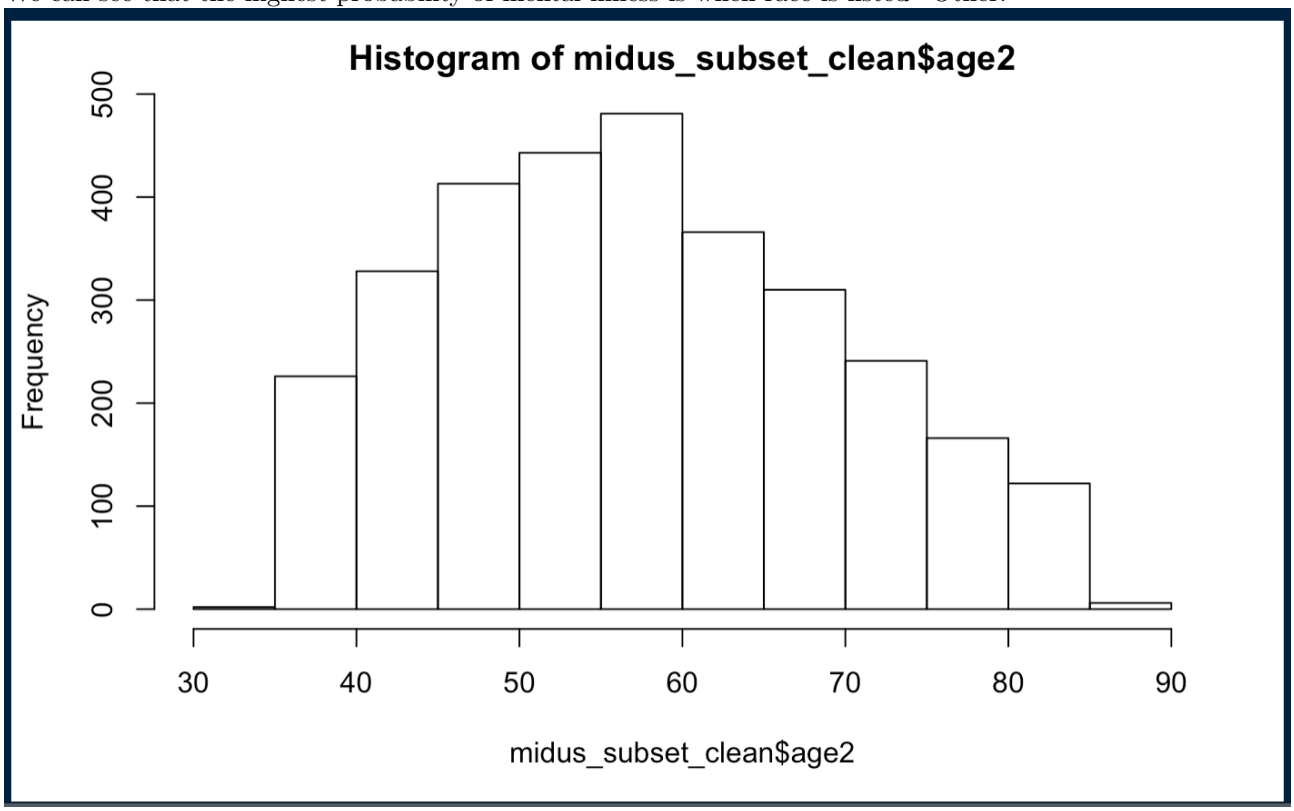
2.0.2 Data Exploration

We can see that the proportion of respondents who report being mentally ill varies based on race of the individual.

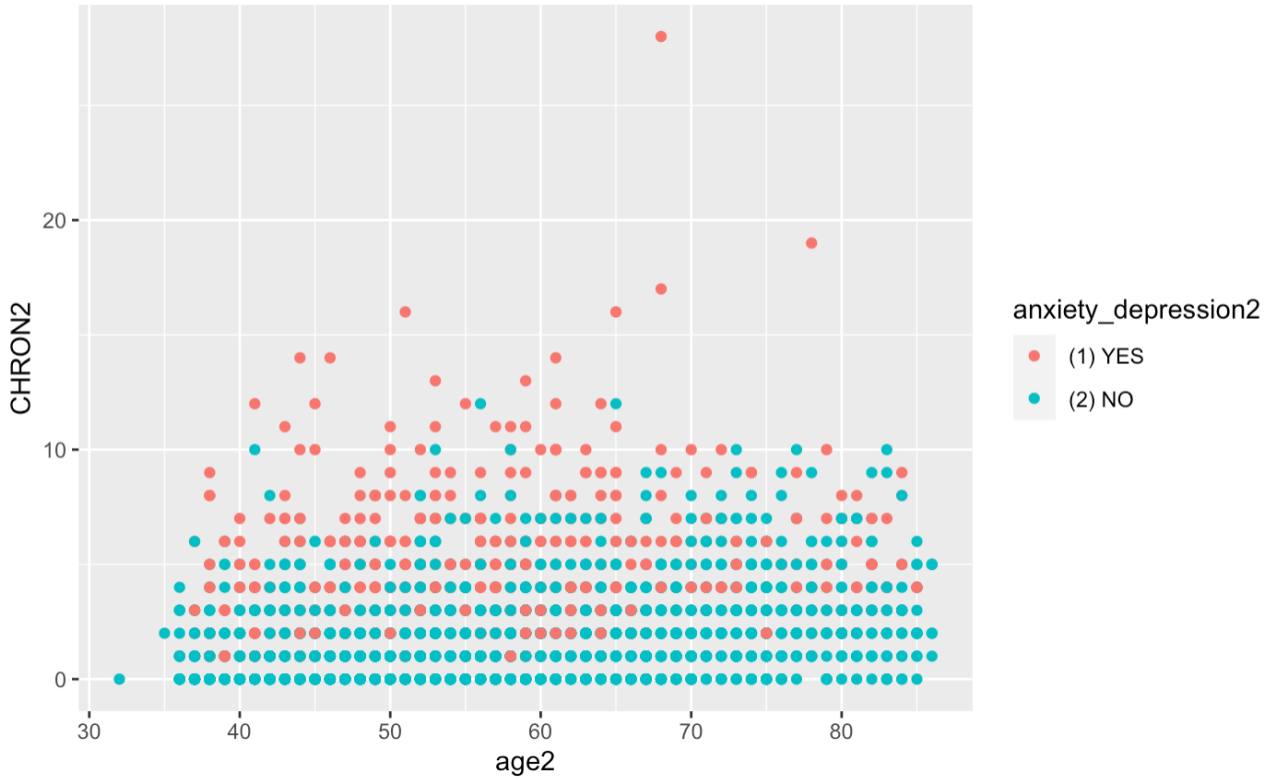
Race <fctr>	probdepression <dbl>
(1) WHITE	0.1853913
(2) BLACK AND/OR AFRICAN AMERICAN	0.1170213
(3) NATIVE AMERICAN OR ALASKA NATIVE ALEUTIAN ISLANDER/ESKIMO	0.1860465
(4) ASIAN	0.1875000
(5) NATIVE HAWAIIAN OR PACIFIC ISLANDER	0.0000000
(6) OTHER (SPECIFY)	0.3181818
(7) DON'T KNOW	0.2000000
(8) REFUSED	0.0000000

8 rows

We can see that the highest probability of mental illness is when race is listed "Other."



We can see that age is roughly normally distributed.



We can see that there is a relationship between an individual's mental health and chronic illness and age.

3 Methods

3.0.1 Decision Trees

The first method that we used to look at possible risk factors for mental illness was classification trees. Tree-based algorithms allow for some stability and ease of interpretation. Tree-based algorithms map nonlinear relationships quite well. This is an adaptable method of problem-solving (classification or regression). The classification tree method is a process that uses a sequential set of questions about x values to classify a prediction on y. These methods involve stratifying or segmenting the predictor space (x-values) into a number of simple regions[4].

Set of X values are which y-hat = 1 has a simple form consisting of a set of rectangular regions. In order to predict a given observation, we typically use the mean of the mode for the training observations in the given class.

We use recursive binary splitting to grow a classification tree. One criterion for making the binary splits is the classification error rate We assign an observation in a region in to the most commonly occurring class The classification error rate, E, is the fraction of the training observations in that region that do not belong to the most common class

$$E = 1 - Max_k(\hat{p}_m k)$$

where p_{mk} represents the proportion of number of observations in the mth region given for a kth class.

If we were just to minimize the classification error rate we would create trees that are not that useful, i.e., we would not gain that much information. We want some kind of trade off between minimizing classification error rate and maximizing information gained. So instead we want to we use a method to measure node purity. These methods measure purity, whether the node contains mostly observations from the same class. [5].

The first method is measuring entropy:

[4]Categorical Data Analysis, by Alan Agresti, Wiley, 2014.

[5]Ch 8: Tree-Based Methods." An Introduction to Statistical Learning: with Applications in R, by Gareth James et al., Springer, 2017, pp. 303–332.

$$D = -\sum_{k=1}^{k} \hat{p}_{mk} log \hat{p}_{mk}$$

Entropy is the measurement of disorder in your data. Entropy is important in measuring information gained by our tree, by measuring the amount of uncertainty from the nodes. We want to minimize entropy.

The next method that is similar is the gini index:

$$G = \sum_{k=1}^{k} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

The Gini Index calculates the probability that "a specific feature that is classified incorrectly when selected randomly." [6]

We use the built in tree() function in r (from the tree package) that uses the Gini Index to originate binary splits.

This is because a tree can continue to grow until there are as many nodes as there are distinct observations in the data, a stopping rule is put into place to allow for some misclassification.

So we need to find away to penalize,λ, for the number of nodes, J.

$$\sum_{k=1}^{k} \hat{p}_{mk}(1 - \hat{p}_{mk}) + [\lambda * J]$$

We can do this by using the Pruning method. The Pruning method is like backward variable selection in regression, building a big tree on the full model and cutting the branches that are not adding to the performance of the tree.

Using the cross-validation error rate we can find the optimal number of nodes.

There are many advantages of using trees for our problem. Trees are easy to understand and intuitive. Some argue that the decision trees more closely mirror human decision-making than than other classification approaches. Trees can be displayed graphically, and are easily interpreted (even better when small). Trees easily work with qualitative predictors, this works really well for our data that has a lot of qualitative predictors. Another advantage is that feature selection happens automatically.

However there are just as many disadvantages to using trees for analyzing this type of data. Trees do not allow the user to add in interaction terms. Trees also do poorly on highly correlated data (because it does not take into account covariance between variables) or repeated measurements. So it is not perfect for analyzing longitudinal data. Another disadvantage to trees is that trees are not robust. This means a small change in data cause a large change in the final estimated tree. Trees also tend to not have as great accuracy as other methods and are often used as only as an exploratory method.

3.0.2 Logistic Regression and Lasso

We can also use logistic regression which is a parametric approach towards classification.

For our problem, we wanted to estimate the probability that someone has a current mental illness (within the last 12 months) given a set of parameters X,

$$Pr(isDepressed = 1|X) \rightarrow p(x)$$

we use a logistic function

$$p(x) = \frac{e^{\beta_0 + \sum \beta_i X}}{1 + e^{\beta_0 + \sum \beta_i X}}$$

In logistic regression we want coefficients that give use values that are close to 1 (high probability for observations are in the class) and values that are close to 0 (low probability for observations are in the class)

So we want to maximize the likelihood function

$$\prod_{i:y_i=1} p(x_i) \prod_{i:y_j=1} (1 - p(x_j))$$

[6] https://medium.com/analytics-steps/understanding-the-gini-index-and-information-gain-in-decision-trees-ab4720518ba8

So we want to find the coefficients that give us the maximized value of the function.

We want to also incorporate shrinkage into our model. Shrinkage is where data values are shrunk towards a central point, like the mean.

To do this we will use the LASSO method because we want to minimize RSS but also have an additional penalty that rewards small (sum of) coefficient values.

$$RSS + \lambda \sum_{j=1} |\beta_j|$$

Because we are only fitting a single model for each λ, we could have many possible values of λ to find the best. To figure out which value is the best we calculate the cross-validation error to make sure the performance is consistent.

The LASSO method assumes that all covariates as continuous, which does not perfectly match with how our data has a lot a factored unordered variables.

3.0.3 Smoothing Splines model

Splines allow us to fit a piecewise degree-d polynomial while requiring that it be continuous.

We want to find a function f(x) that fits the data well. In other words, we want to find the f that shrinks the RSS the best

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

We want to find the optimal degree of the spline. So using leave one out cross validation.

We want our spline to be flexible and not over-fit the data. So our goal is to find a f that makes RSS small but that is also smooth.

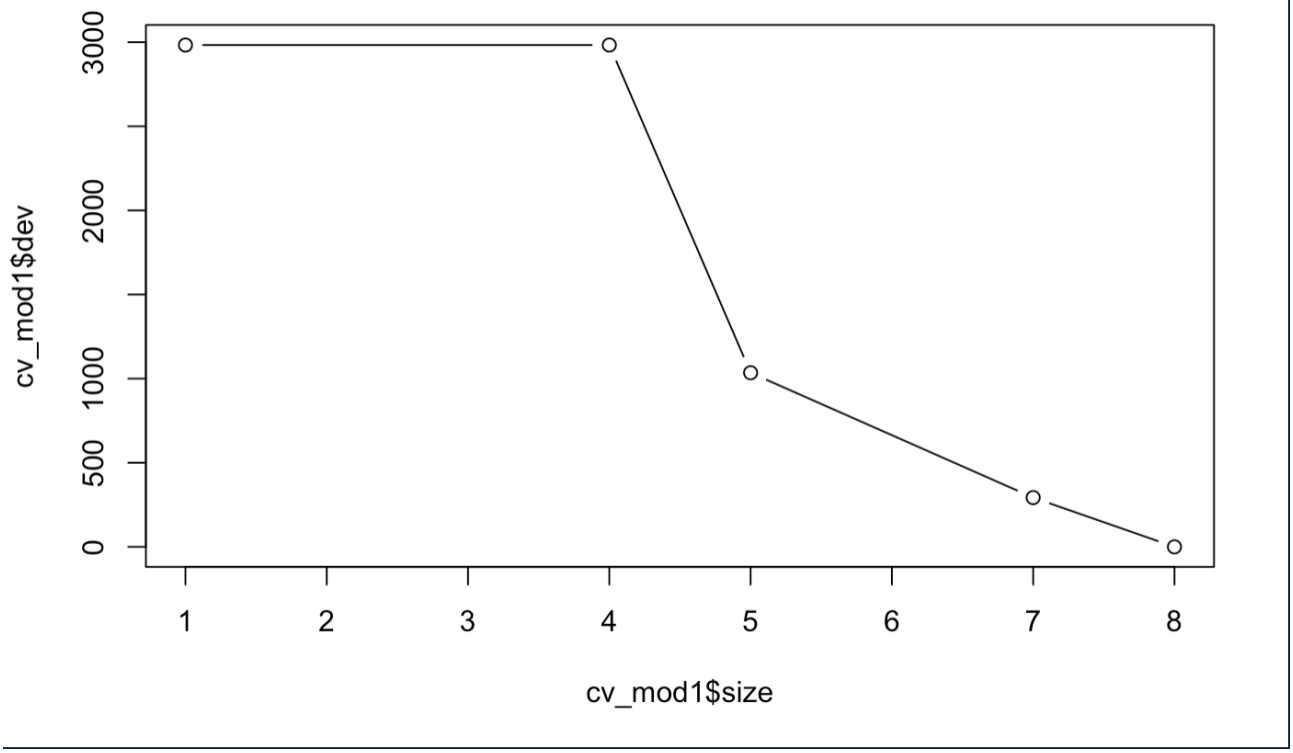
$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int f''(x) dx$$

we are going to use the built in smooth.spline() function in r (from the spline package)

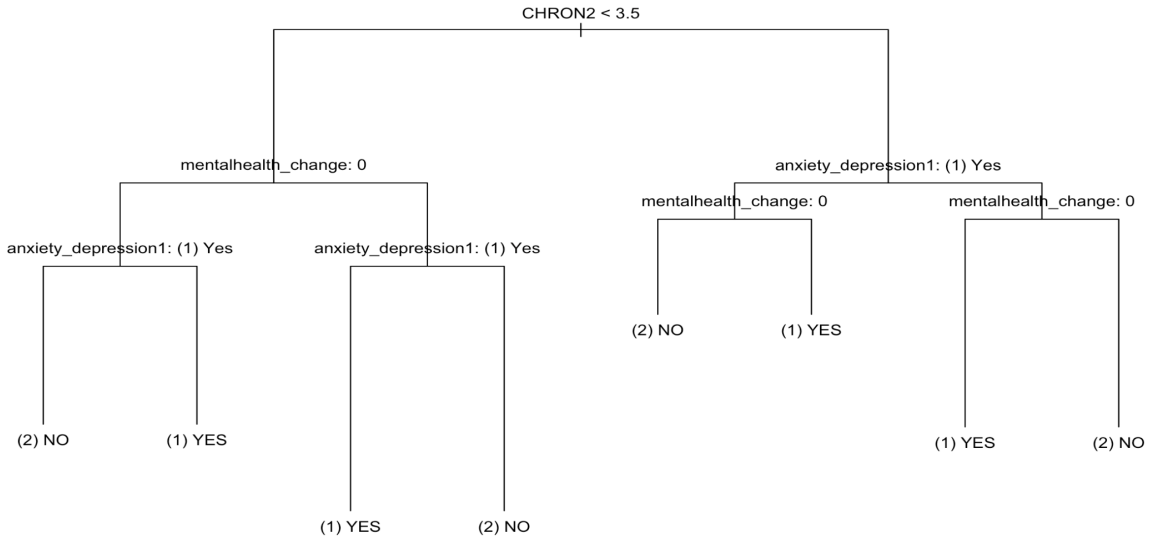
4 Results

4.0.1 Decision Trees

We built a tree in r to predict whether or not the person had a mood disorder in the last 12 months.



We see from this plot that the tree with 8 terminal nodes results in the lowest cross-validation error rate



We can see that most important indicator of current mental health status appears to be whether the number of current chronic illnesses ¡ 3.5. We can also see that whether or not someone's mental health status (in either direction) has changed is also an important predicting current mental health status.

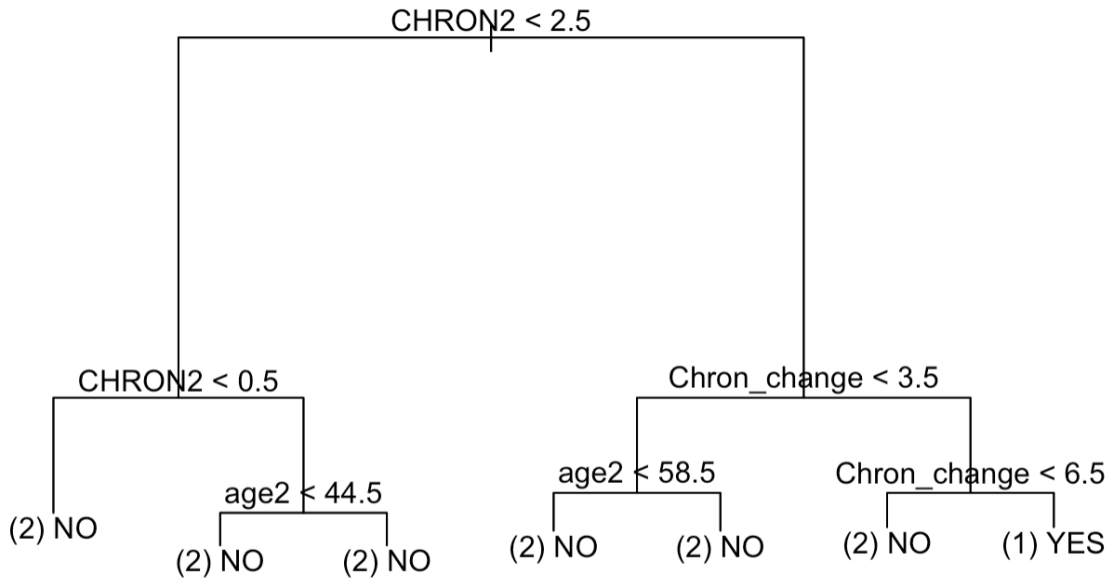
This is not very helpful because we already know that the mental health status of the first wave is correlated to the mental health status during the second wave. So instead we are going to subset on the data.

So to look at the data more closely we are going to break the data into 2 subsets.

first we are going to look at the data set where the individuals did not previously report mental illness

This subset has 2523 observations.

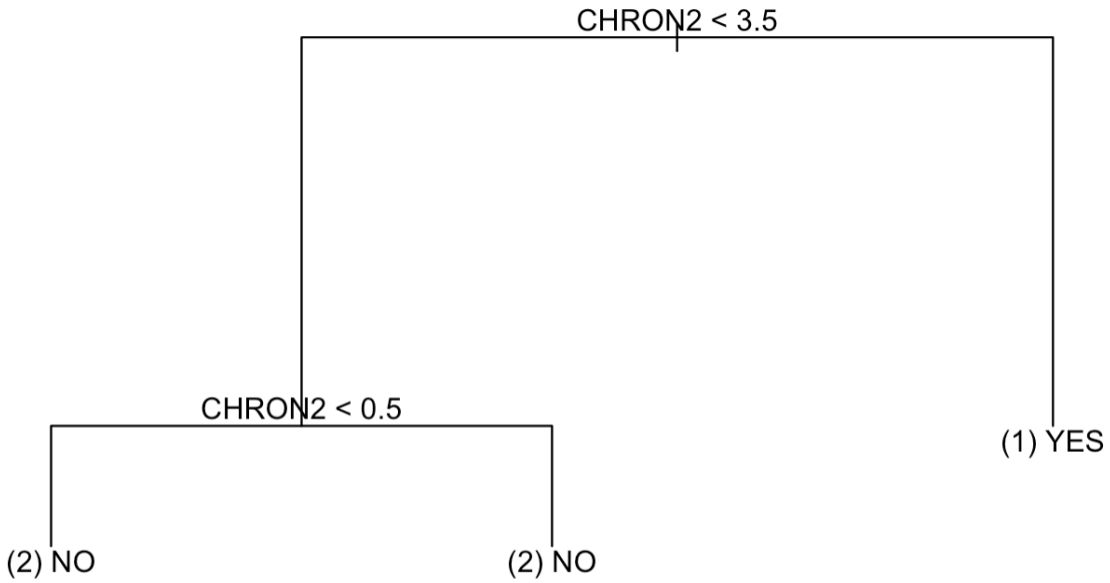
The tree with 8 terminal nodes results in the lowest cross-validation error rate



We can see that most important indicator of current mental health status when no previous mental health illness was reported appears to be whether the number of current chronic illnesses ¡ 2.5. We can also see that age and the increase of the number of chronic diseases is also an important predictor of current mental health status when previous mental illness was not reported.

The next data set we are going to look at is the data set where the individuals did previously report mental illness

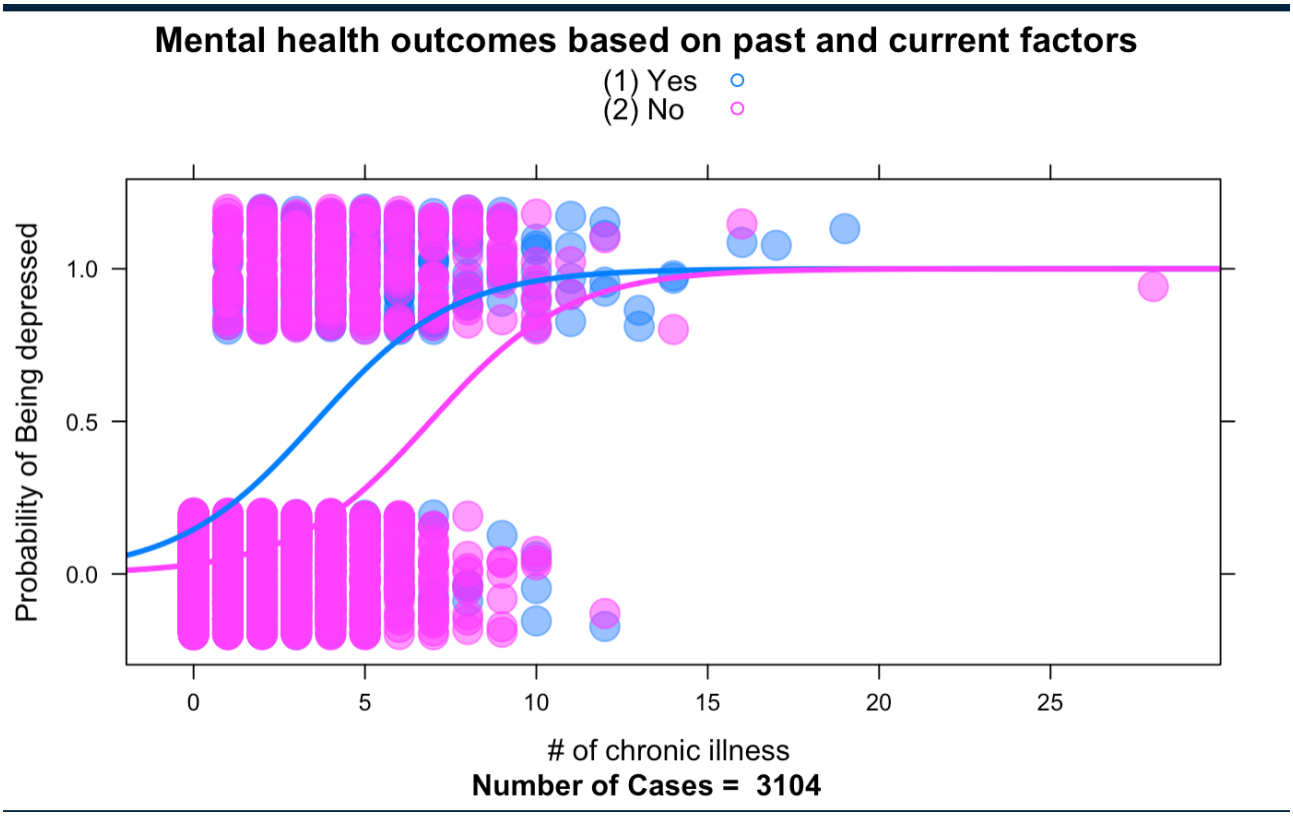
The tree with 3 terminal nodes results in the lowest cross-validation error rate



We can see that most important indicator of current mental health status when previous mental health illness was reported depends on the number of current chronic illnesses.

4.0.2 Logistic Regression

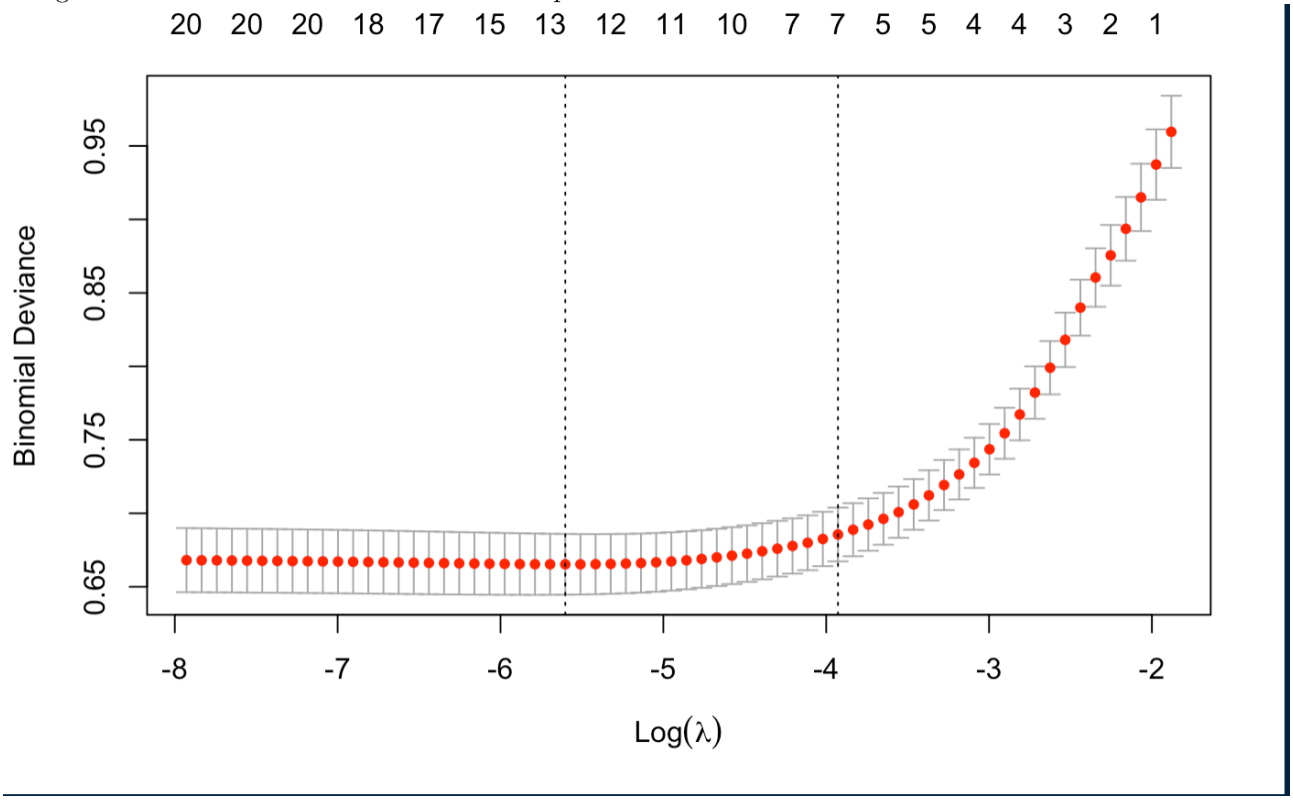
Using what we learned from the decision trees, it seems like the most important variable in predicting current mental health is current chronic illness



Using this plot we can see that as the number of current chronic illness increase so does the the probability of mental illness and that there is a difference between the two subsets.

But we want to use different variables than the one our tree suggests so we going to use. the LASSO method for linear regression.

Using 10 k-fold cross validation we can see the plot of lambda vs cv error rate.



The plot shows that the log of the optimal value of lambda (i.e. the one that minimises the root mean square error) is approximately -5.5.

We are given our optimal λ to be 0.003689419.

We now can see that the the non-zero variables from this LASSO regression are whether or not the the person was depressed, whether or not the person's mental health changed over time, whether someone is Black or other and whether some one's highest level of education changed over time.

```
anxiety_depression1(2) No
                -2.12229384
mentalhealth_change1
                 1.42053442
education_change
                 0.05557767
Race(2) BLACK AND/OR AFRICAN AMERICAN
                -0.76936163
Race(6) OTHER (SPECIFY)
                 0.44226649
```

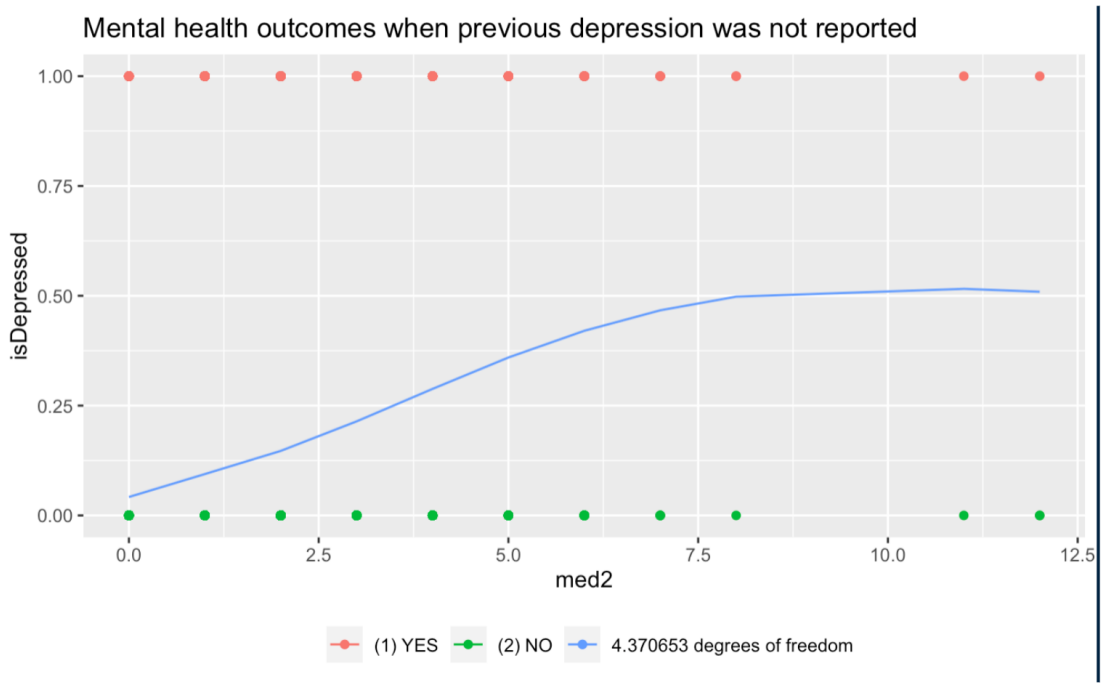
We have an intercept of -32.63245162. The intercept means that for a person who previously reported having mental illness and is not Black or Other and whose education and mental health has not changed, there is 24.6% chance of them having a mood disorder.

We can see that if previous mental illness was not reported the odds of someone having a mood disorder decreases by 2.1229384 (holding all else being constant). We can see that if highest level of education increases the odds (holding all else being constant) having a mood disorder increases by 0.055. We can see that when the race of the person is Black/African American the odds of having a mood disorder decrease by -0.76936163.

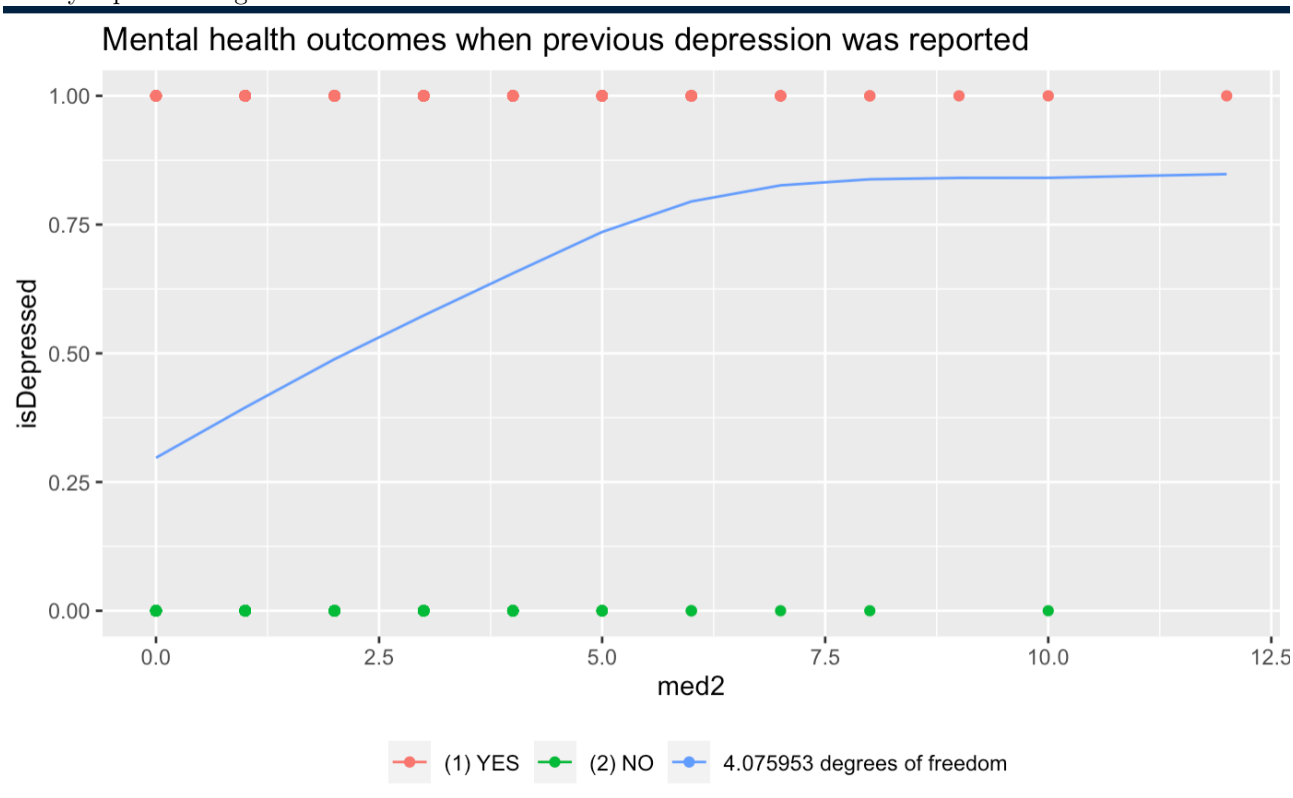
4.0.3 Smoothing Splines model

In order to fit a smoothing spline, we use the smooth.spline() function. And use cross validation to get the degree of freedom and tuning parameter We also split the data into the 2 subsets where the individuals did not previously report mental illness and where they did

We want to find the spline that minimizes RSS the best. So we must first investigate splines by mapping smooth splines to predict probability of current mental health status given that they did not previously report having mental illness using different numeric variables (like number of chronic disease, age, number of current medications someone is on).



We conclude that the one spline that minimize RSS (RSS = 0.3359565) the best is the smooth spline using current number of medication to predict probability of current mental health status given that they did not previously report having mental illness



The one splines that minimize RSS (RSS = 0.7285138) the best is the smooth spline using current number of medications to predict probability of current mental health status given that they did previously report having mental illness

5 Conclusion

We can see that our three methods gave us different results. Using decision trees we are told that an important risk factor for mental illness is the number of current chronic illnesses. Using LASSO we can see that the race of the person, whether their highest level of education changed in the last 10 years and whether their mental health status changed in the last 10 years are all important factors. Using splines we concluded that the number of medications someone is on could be a predictor for mental health status. We also know that there is a lot of multicollinearity within our data, so it is unclear how useful the results from any of these methods are. Although these results appear to be divergent, further exploration of these factors might yield a connection between them. For example, chronic illness may be correlated to an increase in the amount of medication one is taking. Race may also be connected to factors such as chronic illness. Also worth considering are historical events, such as the Iraq War, occurring during the time the data was collected that may have had an impact on the subjects. Another question is whether the stigma of mental illness within certain communities may have affected the validity of the data based on the unwillingness of some of the subjects to answer the questions with complete truthfulness. However, overall, we can conclude that past and current factors in a person's life may play a role in that person's current mental health status.