The Trials and Tribulations of the BRFSS

Justin Weltz and Andrew Brown

5/9/2018

Introduction

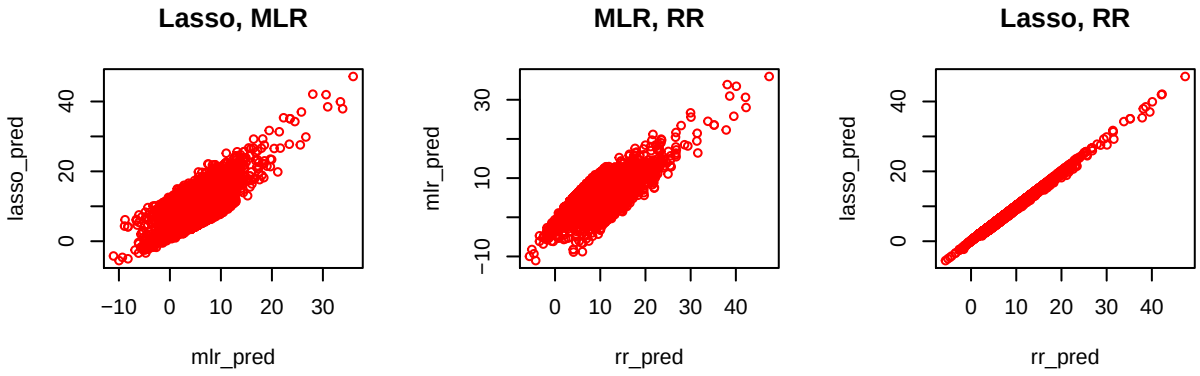
The Behavioral Risk Factor Surveillance System (BRFSS) is conducted by the Centers for Disease Control (CDC) on the United States Population (and is supposed to capture the non-institutionalized adult population older than 18 years residing in the United States). The public data set contains 486,303 observations. Each of these rows is an individual contacted by telephone (this biases the population they are sampling from and may make inferences taken from this study non-applicable to the general US population). There are 279 accessible variables (a lot of demographic information is omitted in order to preserve anonymity) on demographic characteristics, health-related risk behaviors, chronic health conditions, and use of preventative services. However, we will only be studying a subset of these dimensions.

Shrinkage Models

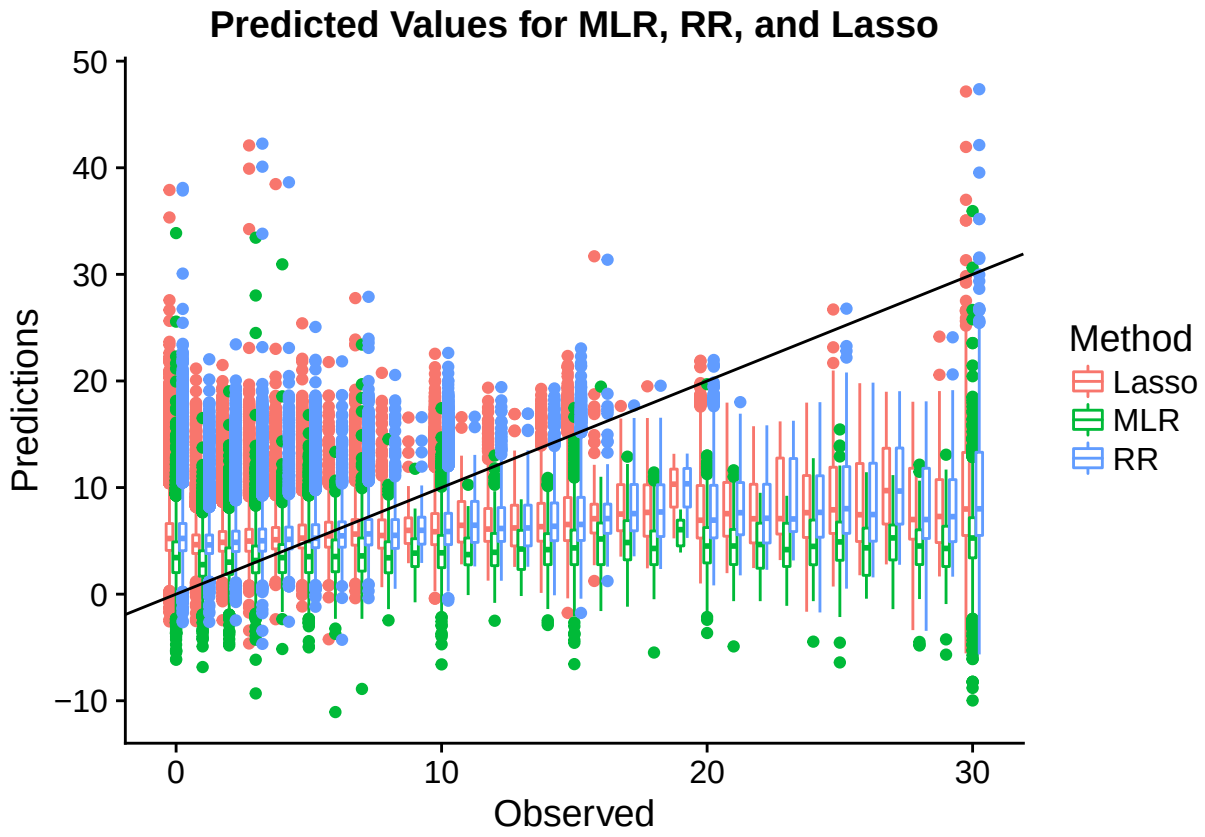
In order to obtain the optimal ridge regression and lasso models, we started by cross validating in order to find the best parameters for these methods. The optimal lambda values for ridge regression and lasso are very small, suggesting the OLS coefficients are good (or relatively good) estimators (as lambda $\to 0$, RR becomes OLS).

Not surprisingly, the ridge regression coefficients were smaller in magnitude than the MLR coefficients (given the constraint on the sum of the squared coefficients). The lasso model also had smaller coefficients and additionally demonstrated some power of variable selection. While the forward and backwards stepwise as well as best subsets variable selection techniques left every variable we proposed (including the interaction variables) in the multiple regression model, Lasso zeroes out the interaction between veterans and average daily drinks! This is interesting because Lasso variable selection is often hypothesized to be equivalent to best subsets (however, clearly they came to different conclusions on our data set!). The general health variable dynamic is also a noteworthy point of comparison between the models. While the multiple linear regression gives new general health the largest coefficient (a weird fact that we discussed at length in the last section), in the RR and Lasso models this variable's coefficient is halved and takes a back seat to sex (the most important variable in the constrained models).

However, looking at the coefficients can only tell us so much about how the model predictions compare. Let's look at some plots:



Before we ran the current models, we experimented with excluding the interaction terms. Since the interaction terms do not have high magnitude coefficients, this did not change the analysis in any substantial way except with respect to the plot above. Without these extra variables, this plot appeared to be a straight line. Essentially, the Lasso and Ridge Regression constraints affected the coefficients in almost identical ways. However, when the interaction terms are added to the model, this line becomes a little smudged (as in the plot above). Ridge regression's spherical boundary (or hyper-spherical boundary depending on the dimension) makes it very hard to completely zero out coefficients. Therefore, only the diamond-like RR boundary can eliminate the interaction between veterans and average daily drinks - demonstrating a fundamental difference between the two models and throwing off the perfect correlation between their predictions.

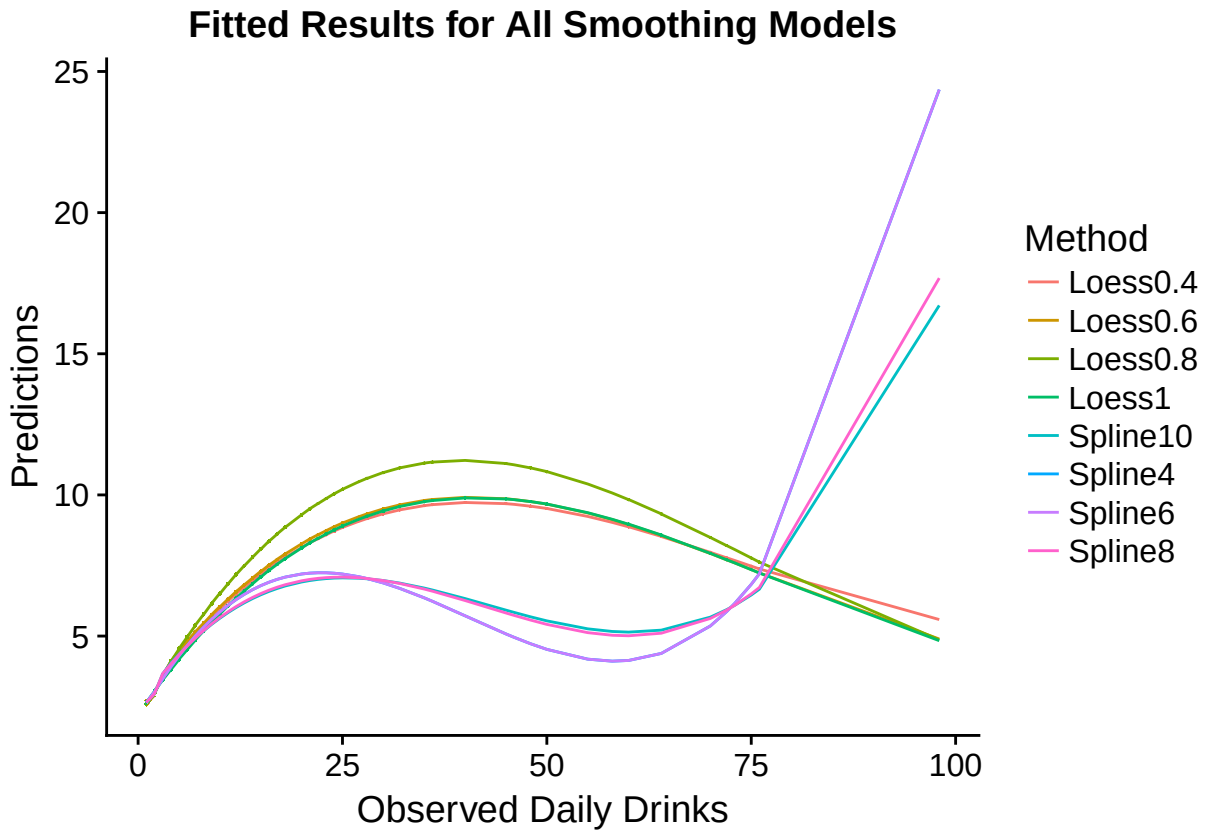


This might be the best way of visualizing the comparison between the three models' effectiveness. For every possible observed value (a discrete scale between 1 and 30 days), we have created a boxplot that depicts the density of predicted values for each method.

First and foremost, we can see that none of these models predict well. The 45-degree line that runs through the graph is a depiction of where mean values should be if the model predicted perfectly. The actual means are nowhere near this standard for a large majority of the observed values. In fact, it looks like all three models are essentially predicting the same value regardless of what they are "supposed to be" estimating. As expected, the Lasso and RR predicted values are very similar to each other - both predict values in the 5-7.5 range for most observations. Interestingly, their mean prediction levels are consistently above mean MLR predictions for every observed value (MLR predictions range from about 4-5). The only indications that larger observed values are having an effect on the models are the higher variance in predictions. It is encouraging to see that the predictions for higher observed values are generally heavily skewed to the right - indicating the presence of many larger predictions in these groups being balanced out by a huge density of low predictions. This finding is consistent with the concern that has been haunting us from the very beginning of this project. The large number of zero observations for mental health days are pulling our predictions (and generally, our model) in weird directions. We will attempt to deal with this issue once and for all in the

sections that follow.

Beyond Linearity



From this plot, we can see that the curves produced by regression splines and Loess actually behave very differently. The regression splines follow a cubic shape, while the Loess models have an inverse parabolic shape. For individuals who report between 15 and 75 daily drinks, the Loess models predict that they will have more days of poor mental health than the regression splines do, while for individuals who report over 75 drinks a day, regression splines predict more days of poor mental health. The predictions for individuals who report over 75 drinks a day should be taken with a grain of salt because the sample size for these explanatory values is not very large. The upward trend in the spline graphs seems to reflect the disproportionate influence of these outlying values, and so does not bode well for the predictive power of this method on the overall data.

Although it is difficult to get a sense of the point density, the general shape of the points does not reflect a functional form (it looks like a sideways parabola, which would fail the vertical line test). This does align with the regression splines' general shape, but could play to the strengths of the loess curves. In fact, we believe that the loess curve does do a better job modeling the data simply because it does not depict a functional form for most of the data. However, distinguishing between the Loess models is much more difficult (since at the end of the day, they're pretty similar). If someone was twisting my arm, I would choose our Loess model with a span of 0.4 to do future predictions on. It has the lowest residual standard error, and it appears to be the smoothest curve, while not predicting extreme values for individuals who report over 75 drinks a day as the regression splines do.

Conclusion

The comparison between the shrinkage and MLR models was very interesting. The change in the relative magnitudes of coefficients implies that we should look more closely into what relationships we can truly

conclude from the data (maybe the conservative analysis that we did in the last section wasn't conservative enough). The plot comparing the different model predictions and observed values was also insightful. The fact that all three models were essentially predicting the same value regardless of the observed response clearly indicates a problem with our predictive power. When looking at non-linear methods, it was interesting to see that the Loess model produced a flat line for the later observations. While a linear model predicts a strong positive correlation between mental health days and average daily drinks throughout the data, maybe the Loess's insistence that there is no relationship in the later range of the explanatory variable is the correct interpretation.

Something New

All our linear attempts at modeling the data have failed, and the huge cluster of observations that report 0 days of poor mental health seem to be a likely culprit for this disaster. Consequently, we decided to change our question slightly in order to get rid of some of these problematic points. The process is simple on paper: We will first attempt to predict whether or not people will have any mental health days. Then, we will only feed the observations we predict to have mental health days into our multiple linear regression.

RANDOM FORESTS

To perform this process, we first decided to turn to a popular machine learning algorithm, Random Forests. In order to describe this technique, it is important to explain how a single tree is created. A binary regression tree is formed by running through each variable in order to pinpoint the best way to partition the observations in a current node into two groups ("best" meaning the division that places the most similar observations in the same group). It is a greedy algorithm that starts with all the observations in the same group and continues to split groups until a constraint is reached or all of the observations in each node (leaves of the tree) correspond to the same response value respectively (they are homogeneous). Unfortunately, individual binary trees (although very easy to understand graphically) are highly variable and overfit the sample. Consequently, by bootstrapping the sample, creating hundreds and hundreds of these trees on these bootstrapped samples, and then averaging their responses for a specific observation we can decrease variance. This collection of trees is a Random Forest.

Although a Random Forest is a non-linear model that performs well on large data sets, it is more or less a black box. It is impossible to infer a functional form from hundreds of independently created binary trees. Consequently, it is difficult to pinpoint the most influential variables and almost impossible to perform any kind of inference. Without a sense for the standard error of test predictions, Random Forest output must always be taken with a grain of salt. When the precision of specific predictions should be prioritized over the general accuracy of the model, the lack of viable Random Forest prediction intervals becomes an issue. For example, while the model may predict the necessary dosage of a dangerous drug for a disease on average, contextualizing model responses with appropriate standard error bars may be the difference between life and death for a specific patient. For this reason, I will attempt to investigate the most natural method that accounts for the variance of points around a model - Root Mean Squared Error.

In a multiple linear regression, Root Mean Squared Error is simply the squared difference between the observed Y value and the response predicted by the model summed together and square rooted. It is a simple formula, but it provides an intuitive sense of how much variability is left unexplained by the model. However, the bootstrapping nature of a random forest adds an interesting element to this concept. Every time a data set is bootstrapped, approximately 1/3 of the original data is not included in the new sample. Consequently, each original data point is independent from the creation of about 1/3 of the trees in the random forest. Predicting every data point on an independently created subsection of the forest gives an unbiased estimate of the accuracy of the model or (when averaged over all the observations) Root Mean Squared Error.

Consequently, in order to create prediction intervals for Random Forests that account for the unbiased variability of points around the model, I will use the concept of out-of-bag RSME. I will also assume a normal distribution of points around the model (although there isn't theory to support this assumption). But, how do we assess these prediction intervals? How do we know whether they succeeded in providing a good assessment of the point variability around the model? Well, if we assume a normal distribution of points around the

model predictions, then 95% of prediction intervals that spans 2*RSME in either direction should capture the correct response variable. Therefore, by getting a sense for the capture rate of my prediction intervals (especially in relation to the prediction intervals of my multiple regression model), I can assess whether they provide a true assessment of prediction variability.

In order to get an accurate sense for the capture rate of these Random Forest prediction intervals I will manually cross-validate using the following procedure:

1. Split the data into ten independent groups.
2. Store these groups as test data sets.
3. Store the training sets (whole data set - test data set) relative to each of the test data sets.
4. Train ten random forests on the training data sets and store them as well as their out-of-bag RMSE.
5. Predict the test sets on all of the corresponding random forest models and create prediction intervals based on the appropriate RMSE.
6. Record the capture rate for the prediction intervals on each of the test data sets.
7. Average the capture rates and report this as the Random Forest prediction interval capture rate.

I performed the same cross-validating procedure for multiple linear regression in order to compare.

Random Forest Prediction Interval Capture Rate:

The crossvalidated RMSE is:

```
## [1] 8.365441
```

The crossvalidated prediction interval capture rate is:

```
## [1] 0.9238103
```

MLR Prediction Interval Capture Rates:

The crossvalidated RMSE is:

```
## [1] 8.487089
```

The crossvalidated prediction interval capture rate is:

```
## [1] 0.9241404
```

The two values above are the cross-validated RMSE and the prediction interval capture rate respectively.

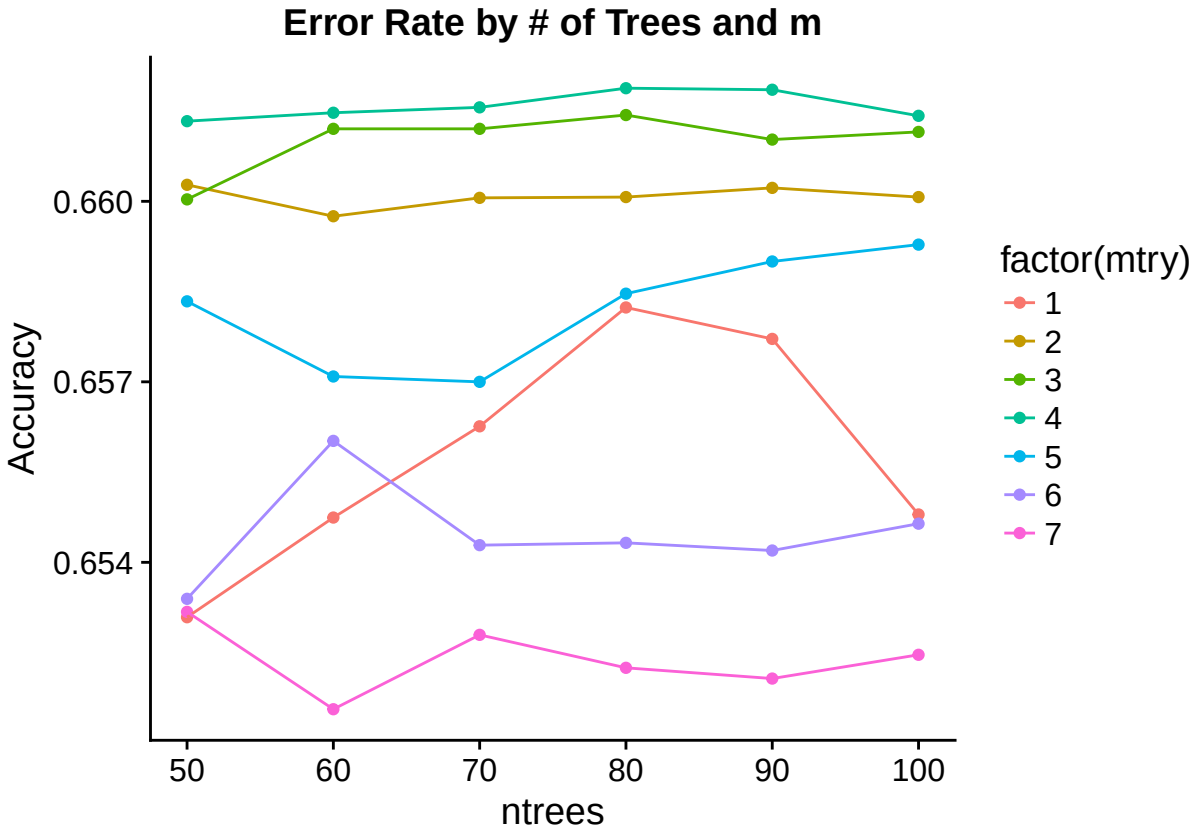
The results are interesting, but not very encouraging (a common theme in this project). First and foremost, it is important to note the immense size of our RMSE for both models. Considering the fact that the response variable is on a discrete scale between 1 and 30 days, an RMSE above 8 indicates that both models would have essentially been better of predicting 15 for every observation (for prediction interval purposes). However, ignoring this reality, it is interesting that both prediction intervals have almost equivalent capture rates. In fact, the prediction interval capture rates are a little higher for Random Forest, while the RMSE for this method is slightly lower. This means that Random Forest predictions are doing better on average. Although Random Forests are an interesting, non-linear approach to modeling the data set, we are still interested in the functional form that a linear model can provide. Consequently, instead of modeling a continuous variable with a regression random forest, we will use a classification random forest and logistic regression to set the stage for a more promising multiple linear regression.

THE TWO STEP APPROACH

Our very first attempts at transforming variables demonstrated that this was going to be a difficult data set to work with. From lopsided residual plots to the mystifyingly positive effect of general health on mental health days, we struggled to glean meaning out of a multiple regression that seemed to consistently defy many of the technical conditions. Although loess and regression splines gave us hope that more flexible methods might be able to describe our 406,000 data points effectively, we were unimpressed with these results as well. It is difficult to pinpoint exactly where our regression techniques are failing so miserably, but we believe that the large imbalance in the data (329,500 people our of 406,000 report that they have experienced zero days of

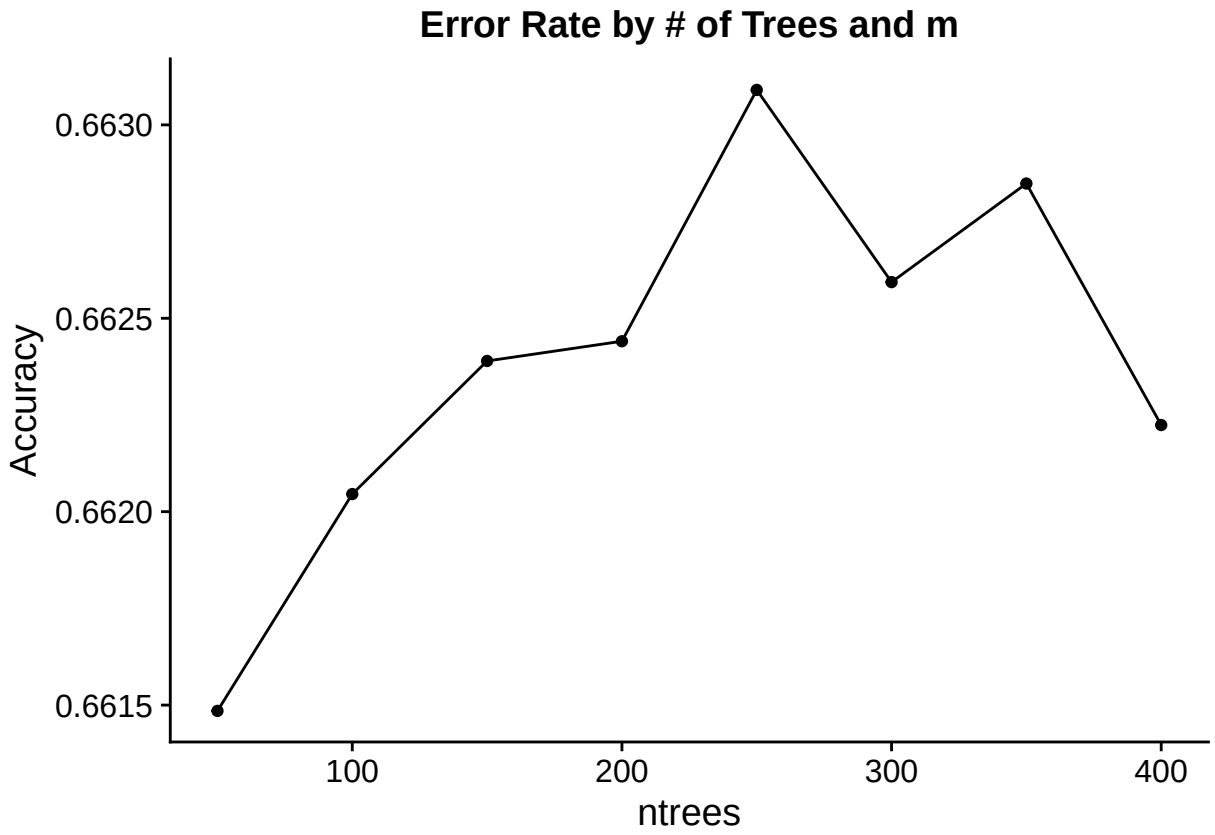
bad mental health in the past 30 days) is causing problems for our models. In order to ameliorate this issue, we plan to divide our regression attempts into two distinct phases. First, we will attempt to predict whether or not an individual has mental health problems (in other words, whether or not their report mental health days > 0). In order to perform this task, we will use two new methods - Classification Random Forests and Logistic Regression. Then, taking our reduced sample size, which hopefully consists of people with non-zero mental health day response variables (although our classification will undoubtedly misclassify some survey participants), and run a multiple linear regression to predict the exact number of mental health days for every individual in this group. Hopefully in this way we will avoid the zero observation's gravitational pull and better predict everyone's' mental health days more accurately.

First, we need to create a binary response variable for our classification algorithms. Then, we will cross-validate to find the optimal parameters for our random forest. We need to determine the best mtry value (the size of the random subset of variables that each split of each random forest can choose from - increases variance in the model to decrease variance in predictions) and ntree value (the number of trees).



The mtry value of 4 clearly performs the best at every tree number.

Next, I will look more closely at the correct number of trees.



250 trees is the optimal number! - although by a small margin (note the y-axis)

Now that we have determined that 100 trees and an mtry of 2 are the best fit for the data set, I will run my final model in order to predict whether a survey participant will report any mental health days or not.

```
##
## Call:
##  randomForest(x = x, y = y, ntree = 250, mtry = param$mtry, importance = FALSE)
##                Type of random forest: classification
##                      Number of trees: 250
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 33.78%
## Confusion matrix:
##      0     1 class.error
## 0 6158 33066  0.84300428
## 1 4803 68088  0.06589291
```

Overall, this is not a great model. The accuracy is not much above 50% and the super high misclassification error rate for those in the 0 category (the ones that don't report any mental health days) means that most observations have been predicted to have mental health problems (predicted as a 1).

In the end, the confusion matrix above tells us that 7312 have been identified as not having mental health issues. This is the number of survey participants that will be eliminated when we run our final MLR model. Sadly, this number is pretty insignificant and we do not expect it will have an important impact on the model. However, we can hope.

Now, let's use logistic regression and compare our prediction error rates. Logistic Regression aims to predict a binary variable through a model obtained is a similar method to MLR. For our case, our binary variable is recorded as a 0 if the individual reports no days of poor mental health in the past month, and a 1 if the individual reports some amount of days of poor mental health.

We start deriving the logistic model by finding coefficients to predict the log odds of a success (p)

$\ln(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \dots$

Then we solve for $\hat{p}$

$\frac{\hat{p}}{1-\hat{p}} = e^{\beta_0+\beta_1*x_1+\beta_2*x_2+\dots}$

$\hat{p} = e^{\beta_0+\beta_1*x_1+\beta_2*x_2+\dots} * (1-\hat{p})$

$\hat{p}(1 + e^{\beta_0+\beta_1*x_1+\beta_2*x_2+\dots}) = e^{\beta_0+\beta_1*x_1+\beta_2*x_2+\dots}$

$\hat{p} = \frac{e^{\beta_0+\beta_1*x_1+\beta_2*x_2+\dots}}{1+e^{\beta_0+\beta_1*x_1+\beta_2*x_2+\dots}}$

This way, we are constraining our probability of success between 0 and 1, and our outcomes predict on a binary variable. This is exactly what we are looking for, so we will try to fit an appropriate logistic model.

For our logistic model, we will split our data into two groups - a training and a test set. We do this to test the accuracy of our model, by building the model based off of the training set, and then testing the accuracy based off the test set. This is an alternative to cross validation, as it assesses the accuracy of the model, but there are no parameters in logistic regression to cross validate on, so we split the data into groups.

The Logistic Coefficients

```
##                (Intercept)                 new_genhlth
##                 0.63506830                -0.05487046
##                    new_sex               new_veteran3
##                 0.75902615                 0.58663053
##               new_sleptim1               new_avedrnk2
##                -0.10376967                 0.09890174
##      new_sex:new_avedrnk2      new_sex:new_veteran3
##                 0.08768556                -0.56109700
## new_veteran3:new_avedrnk2
##                -0.04215902
```

From multiple summaries of our model, we were able to see that there is strong evidence that every explanatory variable is significant in predicting whether or not survey participants have poor mental health days, including the interaction terms!

Interpretation of Coefficients:

Being a female is consistent with an increase in the log odds of reporting poor mental health in the past 30 days of 0.744 in comparison to a male.

Getting one more hour of sleep on average per night is consistent with a decrease in the log odds of reporting poor mental health in the past 30 days of 0.104.

Having one more drink per day on average is consistent with an increase in the log odds of reporting poor mental health in the past 30 days of 0.098.

A 1 unit increase in reported general health (on a scale from 1-5 where 1 is poor and 5 is excellent) is consistent with a decrease in the log odds of reporting poor mental health in the past 30 days of 0.054.

Being a non-veteran is consistent with an increase in the log odds of reporting poor mental health in the past 30 days of 0.58 in comparison to a veteran.

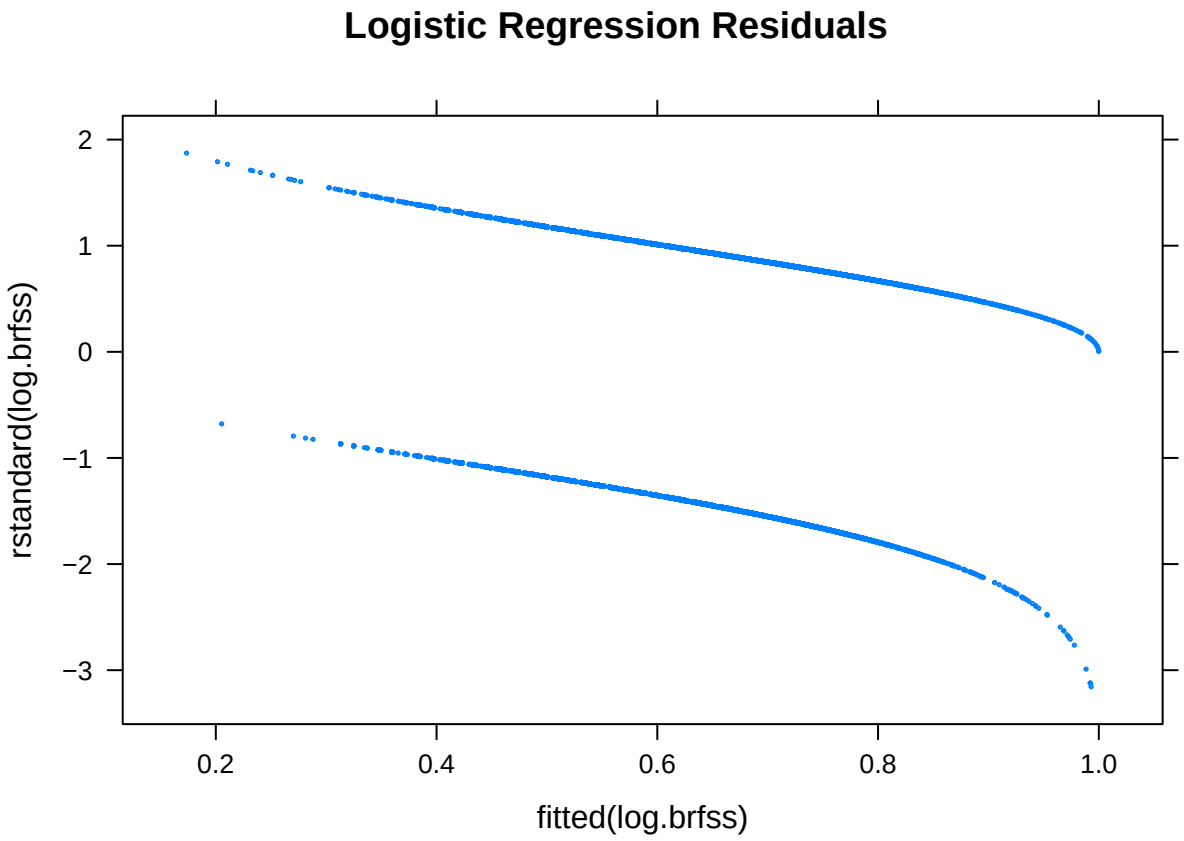
The log odds of reporting some amount of poor mental health days based on drinking increases by a factor of 0.09 for a male in comparison to a female.

The log odds of reporting some amount of poor mental health days based on veteran status decreases by a factor of 0.55 for a male in comparison to a female.

The log odds of reporting some amount of poor mental health days based on drinking decreases by a factor of 0.04 for a veteran in comparison to a non-veteran.

These coefficients are all consistent with our findings from the MLR model, which suggests that the predictions we make based on observations from this logistic model may be more accurate than what we've gotten before, as this logistic model will predict similarly to how our MLR model did in terms of days of poor mental health.

These coefficients are all consistent with our findings from the MLR model, with the exception of general health. The similarities of these coefficients suggest that this logistic model might predict zero vs. nonzero days of poor mental health somewhat accurately. If many observations with zero days of poor mental health are removed, then possibly our predictions from our MLR model will be improved. In fact, the logistic model predicted 18723 observations to have no days of poor mental health in the past month. So while we must always remember that our MLR model was extremely inaccurate and unpredictable, hopefully the absence of these observations will help with our predictions! But first, let's assess the accuracy of this logistic model.



From the residual plot, we can see that the fit is not as horrible as that of the MLR model. The residuals are distributed fairly normally and symmetrically around 0, which is a step in the right direction from our extremely asymmetric residuals we have encountered previously. While these results are better than the residual plots before, and may lead to better results than other models, these residuals are still far from perfect, as the variance is not perfectly constant, and the tail at the end of the plot when the fitted values are 1 make the residuals not perfectly normally distributed around 0. We will further look into the results of our model next, but as of now, this logistic model is looking promising, despite its shortcomings.

Now, in order to see the accuracy of our model, we will try to predict on our test data.

```
## [1] "Accuracy 0.622969360477072"
```

Here, we are predicting new observations from our model. These predictions will return probabilities between 0 and 1, but since we are focused on failures vs. successes (no days of poor mental health reported vs. at least one day of poor mental health reported), we then turn these probabilities into discrete 0 or 1 values determined by if the proportion was greater than 0.5 or not. Once we have this binary vector, we will calculate the number of predicted values that match the observed values (which were turned into a binary 0,1 variable in a similar way) to find the misclassification error rate. One minus this percentage will give us the accuracy of our model, which we can see is around 62%. As each individual observation is distributed on a Bernoulli, there is a 50% chance that we completely randomly predict the observation correctly, so having an error rate that is not much greater than 50% is not very reassuring that we have a good model.

To further examine the accuracy of our predictions, we will now plot an ROC (Receiver Operating Characteristic) curve. An ROC curve is a plot of how well a binary variable is predicted when the cutoff is adjusted. In our case, we will be plotting the True Negative Rate vs. True Positive Rate, as the proportional cutoff for what we consider a success vs. failure varies.

An ROC curve from a very accurate model would be pulled very far up and to the right, so that it comes close to a right angle, creating an area under the curve of close to 1. This is due to the fact that when the true positive rate is very small, you want the true negative rate to be very large, as a small TPR would indicate there are not many successes in the data, and accordingly there are lots of failures, which a good model would predict. As can bee seen from our ROC curve and AUC, this is not the case. This further confirms that our model is not very accurate, as our curve hovers close to the diagonal, and our AUC is approximately 0.606, which is close to the minimum of 0.5 (when the curve is exactly the diagonal). This is not a great model.

The Final Regression

The Original MLR

```
##
## Call:
## lm(formula = actual_y ~ new_genhlth + new_sex + new_veteran3 +
##     new_sleptim1 + new_avedrnk2 + new_sex * new_avedrnk2 + new_sex *
##     new_veteran3 + new_veteran3 * new_avedrnk2, data = brfss.rf.use)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.878  -5.451  -3.060   1.784  35.603
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.89743    0.18231  21.379  < 2e-16 ***
## new_genhlth                 1.72705    0.02539  68.028  < 2e-16 ***
## new_sex                     2.20452    0.24405   9.033  < 2e-16 ***
## new_veteran3                0.91128    0.11817   7.712 1.25e-14 ***
## new_sleptim1               -0.63641    0.01825 -34.876  < 2e-16 ***
## new_avedrnk2                0.34491    0.02594  13.298  < 2e-16 ***
## new_sex:new_avedrnk2        0.20399    0.02345   8.698  < 2e-16 ***
## new_sex:new_veteran3       -1.27607    0.24665  -5.174 2.30e-07 ***
## new_veteran3:new_avedrnk2 -0.06895    0.02899  -2.378   0.0174 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.767 on 112106 degrees of freedom
## Multiple R-squared:  0.06716,    Adjusted R-squared:  0.0671
## F-statistic:  1009 on 8 and 112106 DF,  p-value: < 2.2e-16
```

The MLR based on the Random Forest Predictions

```
##
## Call:
## lm(formula = new_menthlth ~ new_genhlth + new_sex + new_veteran3 +
##     new_sleptim1 + new_avedrnk2 + new_sex * new_avedrnk2 + new_sex *
##     new_veteran3 + new_veteran3 * new_avedrnk2, data = rf_entries.use)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.786  -5.602  -3.092   2.096  34.859
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.20655    0.21124  19.914  < 2e-16 ***
## new_genhlth                 1.84612    0.02728  67.673  < 2e-16 ***
## new_sex                     1.24813    0.27077   4.610 4.04e-06 ***
## new_veteran3                0.12285    0.16743   0.734   0.4631
## new_sleptim1               -0.58921    0.01943 -30.317  < 2e-16 ***
## new_avedrnk2                0.32842    0.03127  10.502  < 2e-16 ***
## new_sex:new_avedrnk2        0.25345    0.02481  10.214  < 2e-16 ***
## new_sex:new_veteran3       -0.50522    0.27278  -1.852   0.0640 .
## new_veteran3:new_avedrnk2 -0.06478    0.03403  -1.904   0.0569 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.831 on 101145 degrees of freedom
## Multiple R-squared:  0.06876,    Adjusted R-squared:  0.06869
## F-statistic: 933.5 on 8 and 101145 DF,  p-value: < 2.2e-16
```

The MLR based on the Logistic Regression Predictions

```
##
## Call:
## lm(formula = new_menthlth ~ new_genhlth + new_sex + new_veteran3 +
##     new_sleptim1 + new_avedrnk2 + new_sex * new_avedrnk2 + new_sex *
##     new_veteran3 + new_veteran3 * new_avedrnk2, data = log_entries.use)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.365  -3.734  -1.846   0.041  31.397
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 5.50189    0.34987  15.726  < 2e-16 ***
## new_genhlth                 1.88714    0.01930  97.763  < 2e-16 ***
## new_sex                    -0.54428    0.33597  -1.620 0.105231
## new_veteran3               -1.56376    0.33970  -4.603 4.16e-06 ***
## new_sleptim1               -0.85808    0.01596 -53.771  < 2e-16 ***
## new_avedrnk2                0.05366    0.02396   2.239 0.025129 *
## new_sex:new_avedrnk2        0.23927    0.01474  16.230  < 2e-16 ***
## new_sex:new_veteran3        1.18267    0.33537   3.526 0.000421 ***
## new_veteran3:new_avedrnk2  0.06732    0.02523   2.669 0.007616 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.049 on 152580 degrees of freedom
```

```
##   (1441 observations deleted due to missingness)
## Multiple R-squared:  0.1008, Adjusted R-squared:  0.1007
## F-statistic:  2137 on 8 and 152580 DF,  p-value: < 2.2e-16
```

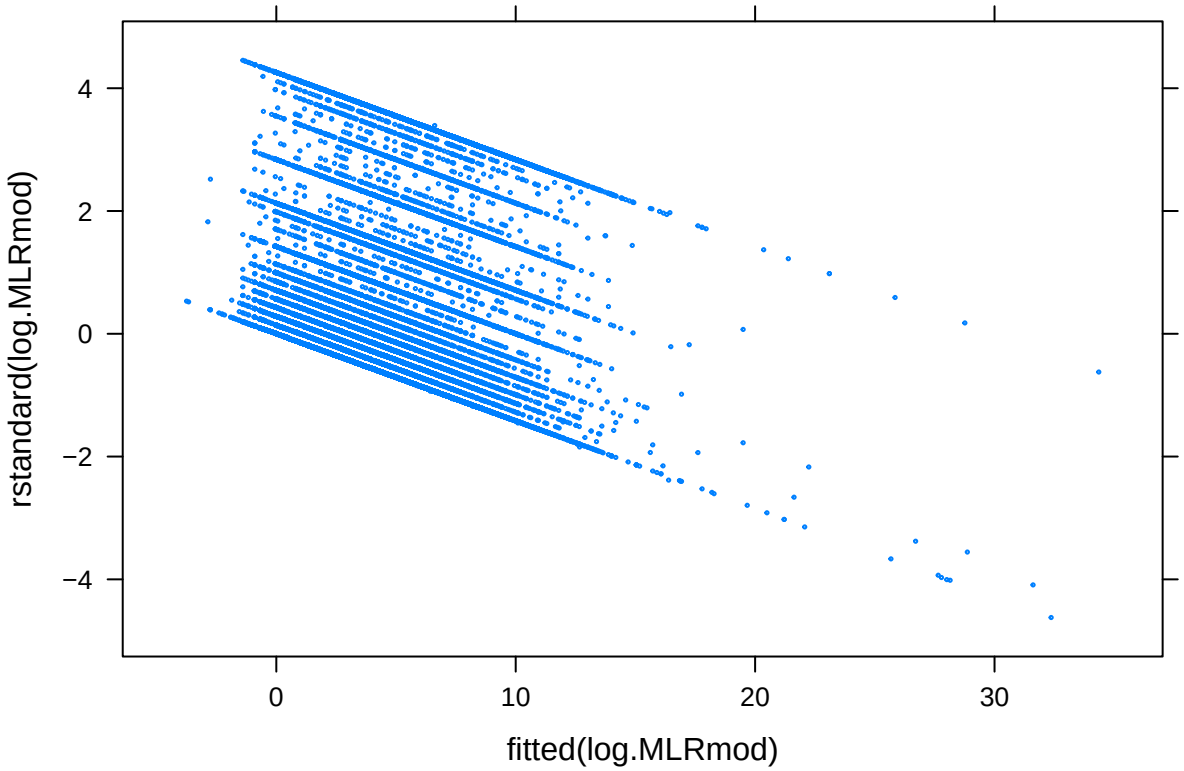
The logistic regression improved the fit of our MLR line! Using the observations that the logistic regression predicted to report some number of days of poor mental health in the past month, we increased our R^2 value from 0.0671 to 0.1011. Now this is still not a very high value for our coefficient of multiple determination, only about 10% of the variation in days of poor mental health is explained by the line. However, this increase seems significant considering that both the original and the random forest MLR are about half of this value.

The different coefficient significances is also a very interesting point of comparison. While the MLR finds that all the variables are significant at the 5% level except for the interaction between veterans and average daily drinks, the Random Forest MLR finds both interaction coefficients to be insignificant. On the other hand, the Logistic MLR actually states that the veteran, average drink interaction variable is significant while the non-interacted daily average drink isn't at the 5% level (this is a red flag since usually the first variables out are the interactions). The Logistic MLR also finds that the sex variable is insignificant even though this was the most important variable in the Ridge Regression and Lasso models, while the Random Forest MLR states that the veteran variable is highly insignificant (a p-value of 0.74!). Clearly, removing some observations from the data using the Logistic and Random Forest Models had some impact on the model! However, the variation between models seems very random and it is difficult to pick up on consistent patterns that might point to deeper relationships.

Summary

As we have mentioned many times throughout this project, this was tough data to work with. While we started off bright eyed and confident in our hypotheses, this optimism deteriorated as we tried one after another transformation, trying desperately to normally distribute our residuals. Sadly, whether it was simply a lack of information or a very odd distribution of points (most of the participants reported no mental health days), we weren't able to satisfy the technical conditions for inference. Consequently, everything we did in the following sections of our analysis attempted to locate small pockets of interest under this shadow. After our multiple regression failed, we turned to non-linear methods. Although random forests seemed like a promising method for partitioning the data, it only performed slightly better. Sadly, it is also very hard to analyze the relationship between variables using a random forest because the concept of inference is not really accessible. We attempted to bridge this gap with prediction intervals which, although they ended up being pretty successful, are only a small piece of the inference that would be necessary to accurately estimate the effect different variables have on the number of mental health days reported. In a last ditch effort, we attempted to decrease the large number of zero responses by using a random forest and a logistic regression to parse the data. In the end, we created two multiple linear regressions that predict the number of mental health days on those predicted to have mental health days in the first place (the first is a quantity between 1 and 30, the second is a binary response). Although this improved our R^2 a tiny bit, it did not change the cursed slant of our residual plot. However, if we had to, we would report this to our client as the final product of our analysis. We tried. In a way, we think the contribution of our work is a narrative of how to attempt and fail to model data in a bunch of different ways. In another sense, this is a cautionary tail demonstrating that, at the end of the day, the quality of the data has to be good or you're doomed.

From the very beginning, we were simply attempting to meet the technical conditions. We tried a series of variable transformations, attempted to remove problematic variables, and compared plots with and without outlying observations. Even the MLR model on the observations that our logistic regression and random forest predicted to be non-zero produced the same slanted residual plot:



It will be forever etched in our memory.

These residuals are horrible! They are nowhere near close to being normally distributed around zero, and the variance ranges from -4 to 4. However, unfortunately, this is the best we were able to get.

Beyond discovering just how difficult our dataset was to work with, the models we ran on it were actually very interesting as well. In our MLR model, we found that general health was extremely significant in predicting poor mental health. In fact, its large positive coefficient demonstrated the opposite of the relationship we would have hypothesized (this means that the number of days of poor mental health increases with general health). We explored this oddity in our MLR section, where we hypothesized that it might be a consequence of the types of people who are likely to report mental health days (the rich and educated are more likely to report and they are more likely to be in good health).

In the Logistic Regression, our inference on general health was actually very different. Although this model is predicting a binary rather than a continuous response variable, it is interesting to see that there are important differences in coefficient values. General health, while still significant, is actually the least significant non-interaction term in the model, as it returns the greatest p-value from the t-test for significance. This result actually compliments the ridge regression and lasso models, in which the general health variable also is shown to be much less important. And, even more interestingly, in the Logistic Regression the coefficient is now negative, which follows the logical hypothesis that better general health would lead to better mental health (less days of poor mental health).

The Random Forest Prediction Intervals were remarkable in that they actually sort of worked! For once in our analysis, we achieved our stated goal. The cross-validated intervals captured 92% of observations (which is pretty close to 95%). It was also surprising to see that the prediction intervals created based on out-of-bag RMSE captured slightly more than the intervals we created in our MLR. The fact that the RMSE for the random forest was slightly less while the prediction intervals captured slightly more points seems to indicate that the Random Forest model was marginally more accurate overall.

It is interesting that the MLR generated from our logistic regression predictions generated the best overall fit. Although, it is difficult to glean any significance from this finding since there doesn't seem to be a consistent story (in terms of variables becoming more or less significant) transitioning from the old MLR model to the new ones. At the end of the day, I am still really not sure which variables are important in determining mental health days. It all seems very random.

As far as the methods we used throughout our analysis, we mostly chose them out of necessity. When running the Multiple Linear Regression, our residual plots were so bad (they were similar to the one above) that we needed to be extremely conservative in our analysis of the coefficients. Instead of being able to interpret the coefficients as they were, we had to generalize them to their sign and relative magnitude (large vs. small) because we could not contextualize our estimates with accurate standard errors. Finally, since the vast majority of our observations were survey participants with no days of poor mental health, we had to find a method of combating this cluster's large influencing over our MLR model. We decided to try predicting binary values for mental health, and only including the observations that were predicted to be non-zero in the final MLR models. To create these binary predictions, we used random forests and a logistic regression model. After running our MLR on these predicted observations, we got a better fit from the logistic MLR, but not much better. This improvement may be interpreted to confirm our theory that the zero observations were having a negative influence over the MLR's predictive power. However, the vast differences between the random forest and logistic MLR's output, in terms of fit and coefficient estimates, discourage any confident conclusions. Generally, it seems that the quantity of these zero observations and the complexity of our data was too much to overcome.

Overall, our data proved very difficult to work with. Although we had such a huge sample to work with (486,303 observations) at first glance, this initial number is deceiving. There were about 375,000 observations that had an NA in at least one of the variables we included in our final model; that's more than 3/4 of our entire sample. At the end of the day, we don't really know if this was random exclusion, or if there was something systematic to the NA's that biased our model. It is also important to note the types of variables we were working with.

Furthermore, it is necessary to acknowledge the inherent problems with the survey's attempt to measure participant's mental health. What constitutes a day of poor mental health? Feeling sad? Having a hallucination? Needing medication? There is no set definition of poor mental health widely known by the public. And, in addition to the responses being subjective, mental health also carries a negative stigma, meaning many people may not be willing to admit that they experience poor mental health. Take a veteran - somebody who has been exposed to numerous traumatic experiences and is likely to suffer from poor mental health, but also someone who has been surrounded by a culture of "the weak don't survive" - they probably aren't very likely to report poor mental health to some random researcher over the phone. This brings us to another potentially problematic point: what type of person is likely to have a landline in this day and age? Additionally, who volunteers intensely personal information to a complete stranger? A landline is more likely to be owned by someone who is older, and possibly from a rural area (as there is less constant cell reception). These types of people may be less aware and accepting of mental health issues, so they would be less likely to report poor mental health (possibly resulting in our large number of zero observations). While it is difficult to find another way to gather important personal information on a large range of people in the United States, the nature of the BRFSS survey most likely biases our data and makes it very difficult to make conclusions that are generalizable to the American population.

Looking back on all the analysis we did on this data, we were pretty much doomed from the beginning. While the five variables we chose seemed to be by far and away the best choices for explanatory variables at the beginning of our analysis, in hindsight, it wouldn't have hurt to look at more. Our model may have been underspecified from the very start - there might have been a variable hiding in the data set that could have rectified our residual plot. We also hypothesized, from our dot plots in the baby stages, that there may not have really been a linear relationship between our explanatory variables and days of mental health. This may have been a problem that should have been approached from the very onset by non-linear methods like random forests and support vector machines.

While there are many things we can look back on and second guess - other ways we could have conducted

our analysis - it is important to keep in mind that mental health has been studied for decades by some of the brightest minds on the planet, and there are still lots of questions left unanswered. Mental health is a complex issue, and while it is a worthwhile topic to look into, we probably were not going to make any groundbreaking discoveries from the BRFSS survey. We tried, but maybe the topic has to be better solidified in the public consciousness before a large-scale survey can generate meaningful results.

Sources

"Behavioral Risk Factor Surveillance System." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 22 Mar. 2018, www.cdc.gov/brfss/index.html.