S670 Final Project

The Fantastic Four
Erik "Human Torch" Parker
Emily "Invisible Woman" Rudman
Vinay "The Thing" Vernekar
Jervis "Mister Fantastic" Wang

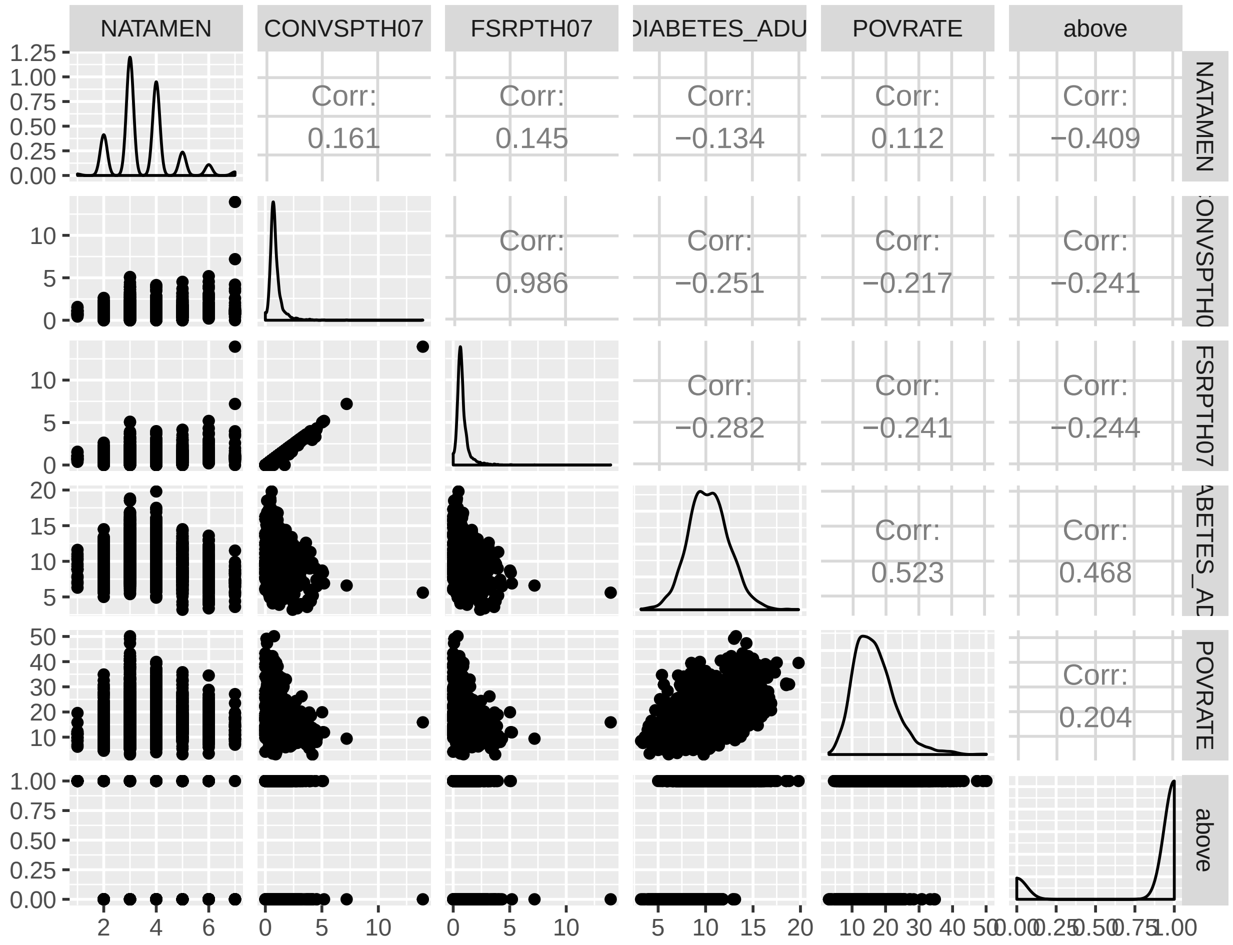
April 11, 2017

Background and Goals

Obesity is one of the leading causes of death in both the US and around the world, and this newly classified disease is viewed by many organizations as one of the most serious public health crises of this century. A more complete understanding of the relationships between various socioeconomic, food environment, geographic, and local community characteristic factors and the rate of obesity may allow public health workers to better address this epidemic in the United States. To meet this lofty goal, the Fantastic Four chose to answer the following two questions. First, what factors, from the limited subset available to us, are most strongly correlated with obesity rate in US adults? And secondly, are we able to reliably predict the probability of populations showing a greater than average obesity rate based just on the factors identified previously? It is our hope that the answers to these two questions will illustrate novel approaches to the prevention, and treatment of obesity in the United States.

Description of Data

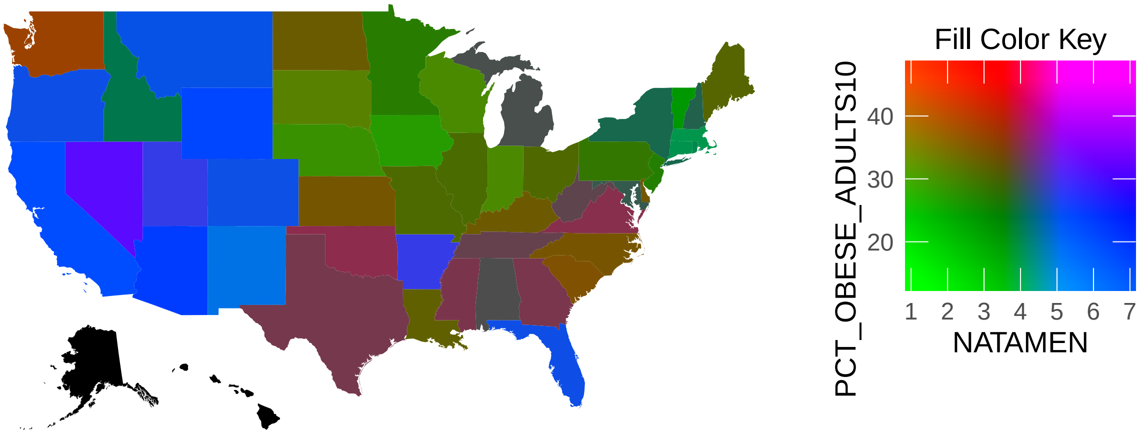
The data for our project were obtained from the long-running "Food Environment Atlas" project from the US Department of Agriculture. This project keeps track of a wide range of variables including food environment factors (such as store and restaurant number and type, local food availability, and poverty assistance programs), as well as more general information such as socioeconomic characteristics, and disease rates for a number of years - all on the county level. The data from this project is quite extensive, with at least some variables reported for all counties in the United States. As our goal was to investigate how different factors impact obesity, we paid particular attention to the obesity rate variable included in this dataset which was reported by county. The sampling and data collection methodology of this project are not reported, but as all of the variables included in this dataset are health, environment, business, or socioeconomic in nature it can likely safely be assumed that they were collected by the Department of Agriculture from public records or in collaboration with other governmental agencies, such as the Department of Health and Human Services and the Department of Commerce.



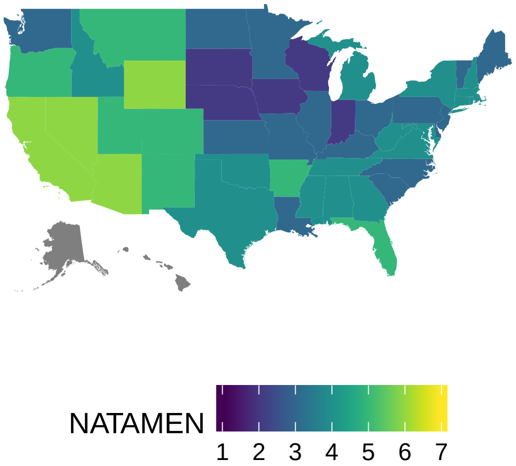
We started by sorting through variables to determine which ones we would want to look at for our project. After choosing about 20 relevant variables, we removed the NA values from our data set.. After we did this we created a variable called above which assigned a 1 to a county if it has a PCT_OBESE_ADULTS greater than the national average and a 0 otherwise. We then looked at the ggpairs plot of those variables with our new above variable. Out of all the variables NATAMEN, CONVSPTH07, FSRPTH07, POVRATE, and PCT_DIABETES_ADULTS10 had the strongest correlation values with above. It is already well known that diabetes and obesity are strongly correlated, so we decided to see if we could fit a model without this as a predictive variable to find more interesting results. Additionally, we see a very strong correlation between the number of convenience stores and full service resturants, and so only need to include one of the two in our final model.

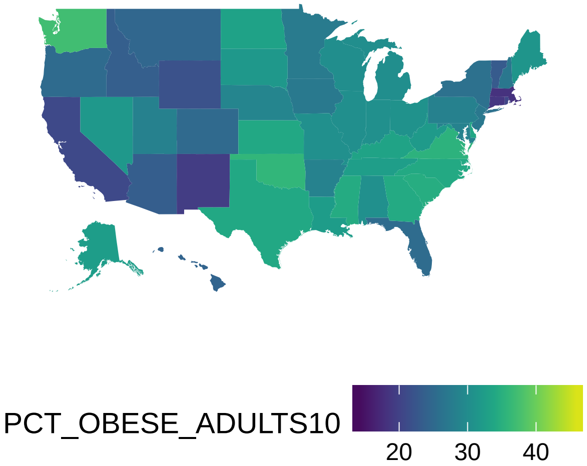
NATAMEN, or the Natural Amenities scale, is an index developed by the USDA in 1999 which ranges from 1 to 7 and measures the desirability of a particular location (here a county) based on natural factors. The physical characteristics chosen to generate this index are: warm winter, winter sun, temperate summer, low summer humidity, topographic variation, and water area. More "desirable" locations on this scale are assigned values closer to 7. CONVSPTH07 is a measure of the number of convenience stores in a county, per 1000 residents, in 2007, and the similar FSRPTH07 is the same measure but for full service restaurants. These two were given for both 2007 and 2012, but only 2007 was chosen as we were working to predict obesity rates in 2010 and so needed measures from preceding years. Finally, POVRATE is the average poverty rate of each county in 2010.

To explore our data further visually, and look beyond correlations, we generated a series of plots and maps.



This first map shows how both NATAMEN and obesity rate vary together on a state level. This map is interesting, as it shows that in general there seems to be a regional relationship between both obesity rates and NATAMEN scores. Also interestingly, though likely not unexpectedly, high NATAMEN is generally matched by low obesity rates and vice versa. To more clearly examine whtat is going one with these two variables we generated two further maps with NATAMEN and obesity rate alone.





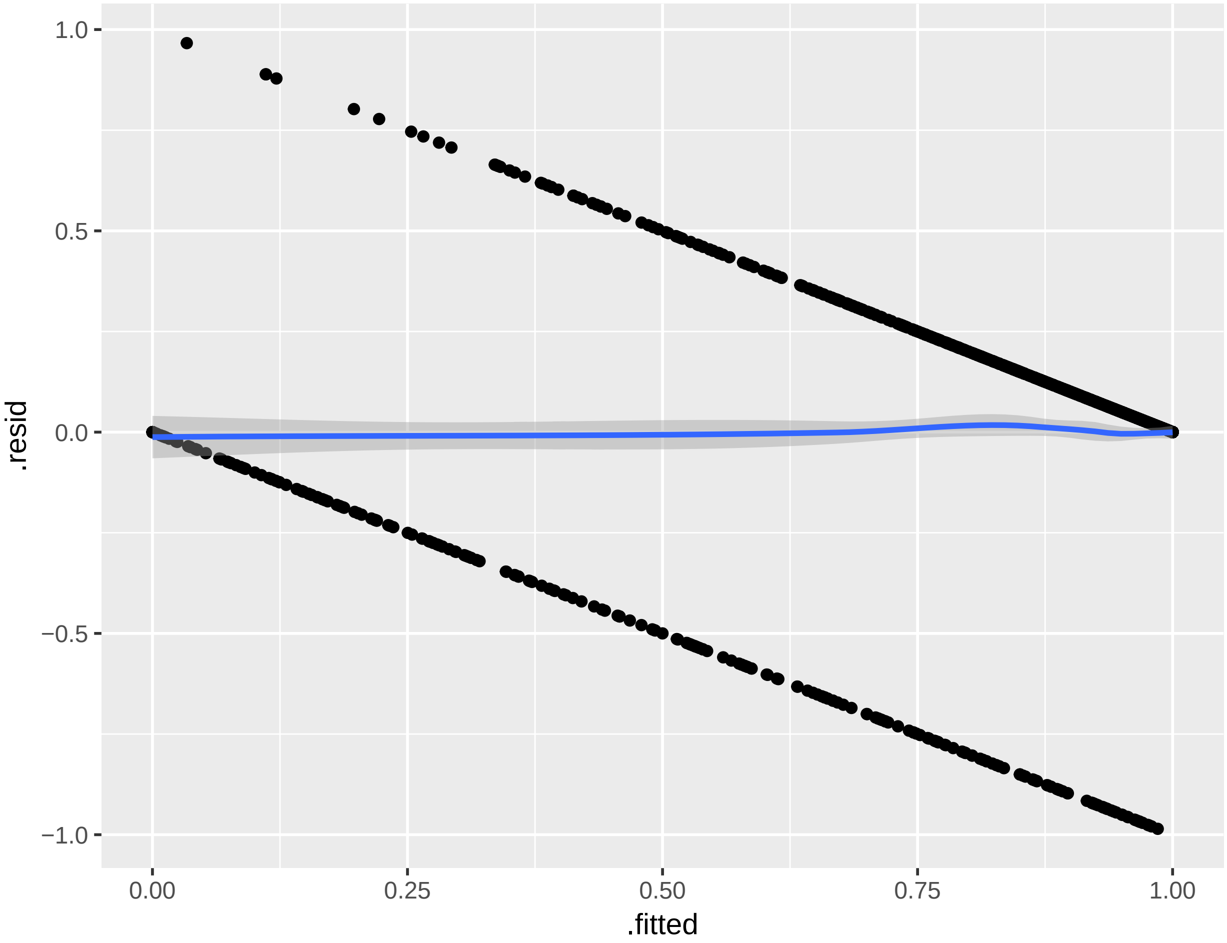
Like the above map, we see from these two that NATAMEN scores are generally clustered by region, with the highest scores seen in Western states and the lowest in midwestern ones. Likewise, obesity rates are also generally clustered regionally - with lower rates generally in the West and higher ones in the midwest and south. These maps, while intuitive, do not explain everything going on though and there are clear outliers, such as the high rate of obesity seen in Washington state despite its high NATAMEN score, and the extremely high rates of obesity seen in the south despite the regions relatively high average NATAMEN scores.

Answering Questions

The previous plots and maps helped us answer our first question at issue: what factors, from the limited subset available to us, are most strongly correlated with obesity rate in US adults? We saw that the factors most strongly related to obesity rate in this dataset are NATAMEN score (and related state information), diabetes rate, poverty rate, and the number of full service resturants and convenience stores per 1000 people. Of these variables, we chose in the end to incorporate NATAMEN, poverty rate, and full service resturants as predictors. Finding the answer to our first question largely served the purpose of allowing us to build the best model possible to answer our second, larger question as identified above.

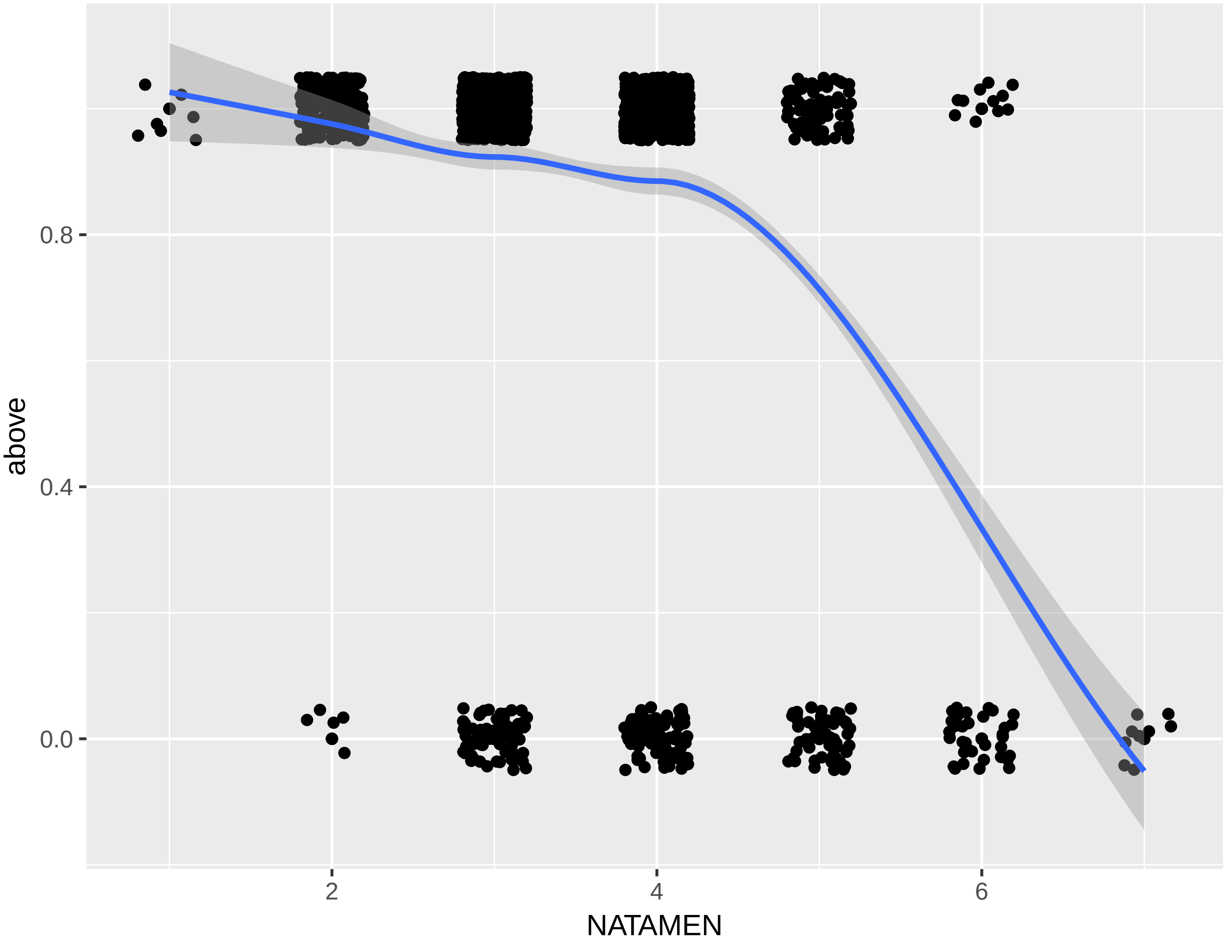
When setting out to construct our model, we decided we wanted to build one with some predictive power. This was initially complicated by the fact that our main response variable was a rate, and so any predictions made would be difficult to interpret. To get around this problem, the team decided it would be best to build a binomial model, and derive predictions from our response being either above or below some defined value. The value chosen for this binomial response was the average, country-wide, obesity rate in 2010 - which ended up as 26.7%. As noted above, with this value chosen, we created a new variable, "above", which denoted whether or not any particular county had an obesity rate above the national average in 2010. This new response variable finally allowed us to accomplish our goal and generate a predictive, easily understandable, binomial model giving the probability of a county being above the national average for obesity based on our three chosen explanatory variables.

```r
model.ob <- glm(above ~ FSRPTH07 + NATAMEN + POVRATE10, family = "binomial",
    data = model.data)
# Binomial model predicting if counties are above average obesity from full
# service resturants/1000 people, NATAMEN, and poverty rate.
```



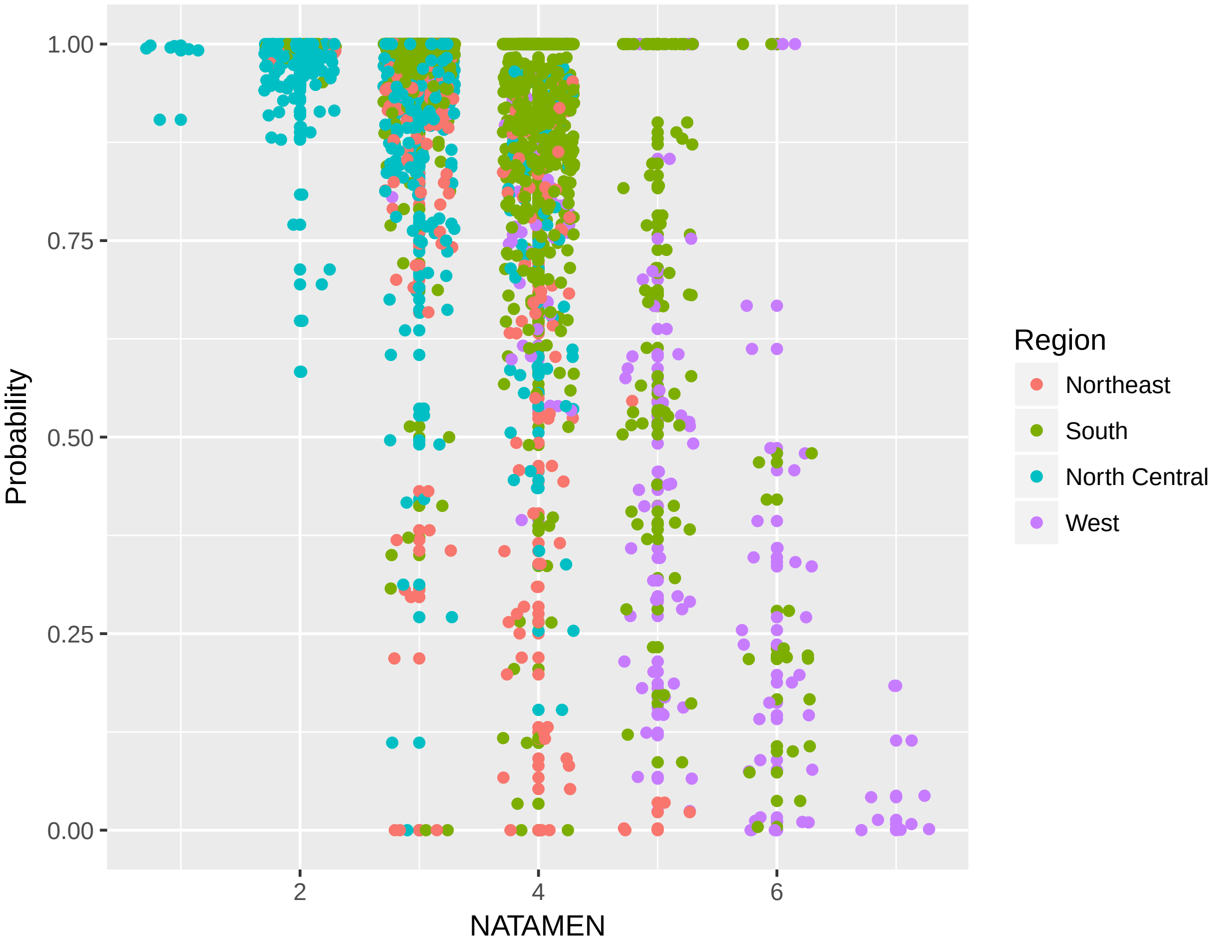
Overall, we were very happy with the model we were able to fit with only three predictive variables. We fit many variations of this model, including all possible interactions. None of these other alternative models lead to a significantly better fit or rediction in deviance though, so we decided to go for the most simple additive model possible. Another benefit of this fully additive model is its ease of interpretation - all effects seen here come directly from the influence of our three variables individually, not from their interactions. However, one of the main limitations we faced which may impact the accuracy of our model was an incomplete data set. Some of the counties had missing information in one of the variables we studied, so we made the choice to remove those values, thus allowing the model to be constructed, but losing some of our original explanatory power.

Before exploring the predictions generated from our model, we decided to quickly look at our raw model data - in terms of counties above and below the national average for obesity.

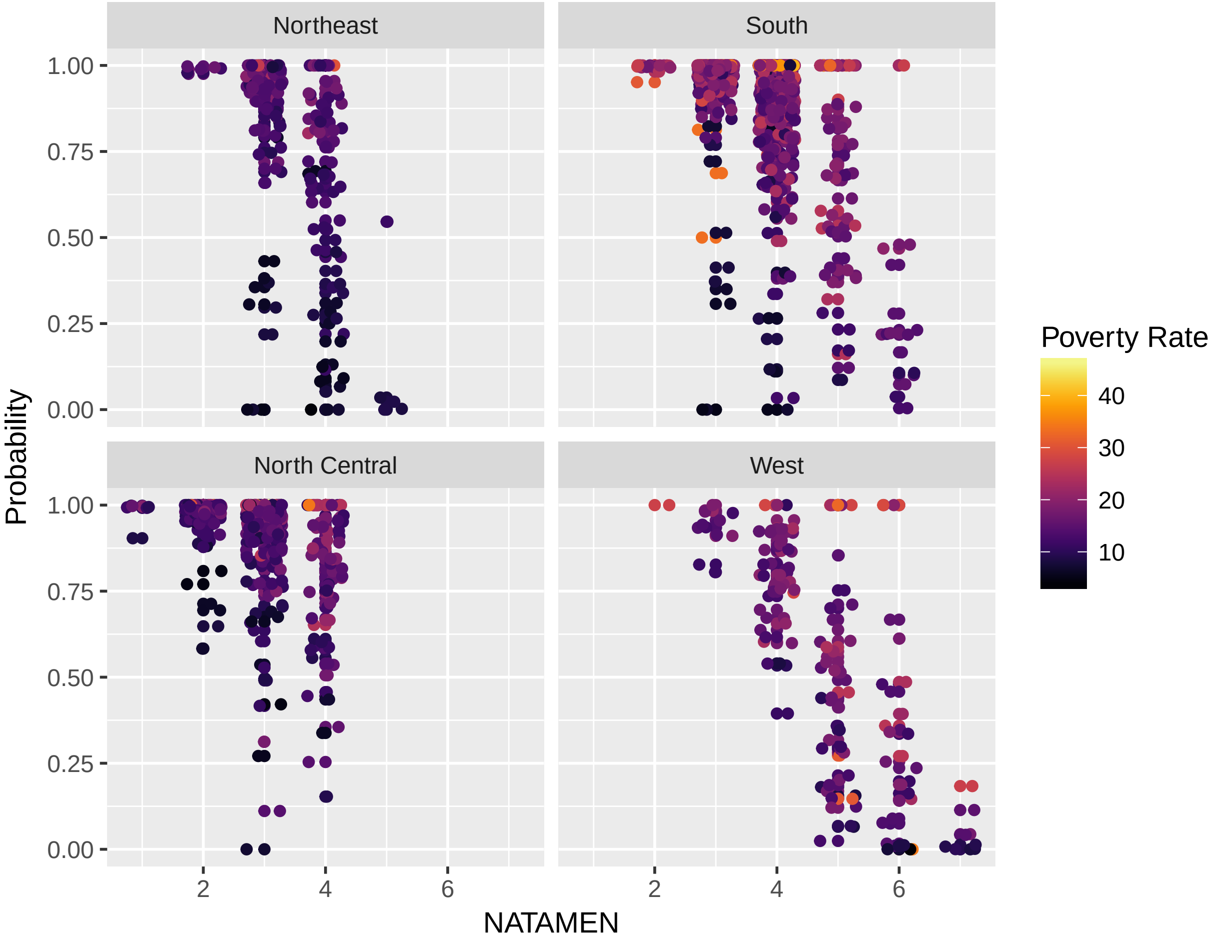


The above plot shows that there is a pretty clear relationship between counties being above/below the national obesity average and NATAMEN scores. No counties with a score of 1 (the lowest) are seen to be below average, and no counties with a score of 7 (the highest) are above average. Beyond that, we see that until we reach a NATAMEN score of 5, the majority of counties are above average, and after this point, there is a quite sudden switch with counties largely being below average for obesity rate. This plot aslo shows us, interestingly, that the majorty of counties included in this analysis are above average for obesity. This likely means that many counties are just above the national average of 26.7, and any counties far above the average must be counteracted by counties far below the average - in order to keep the national average where it is.

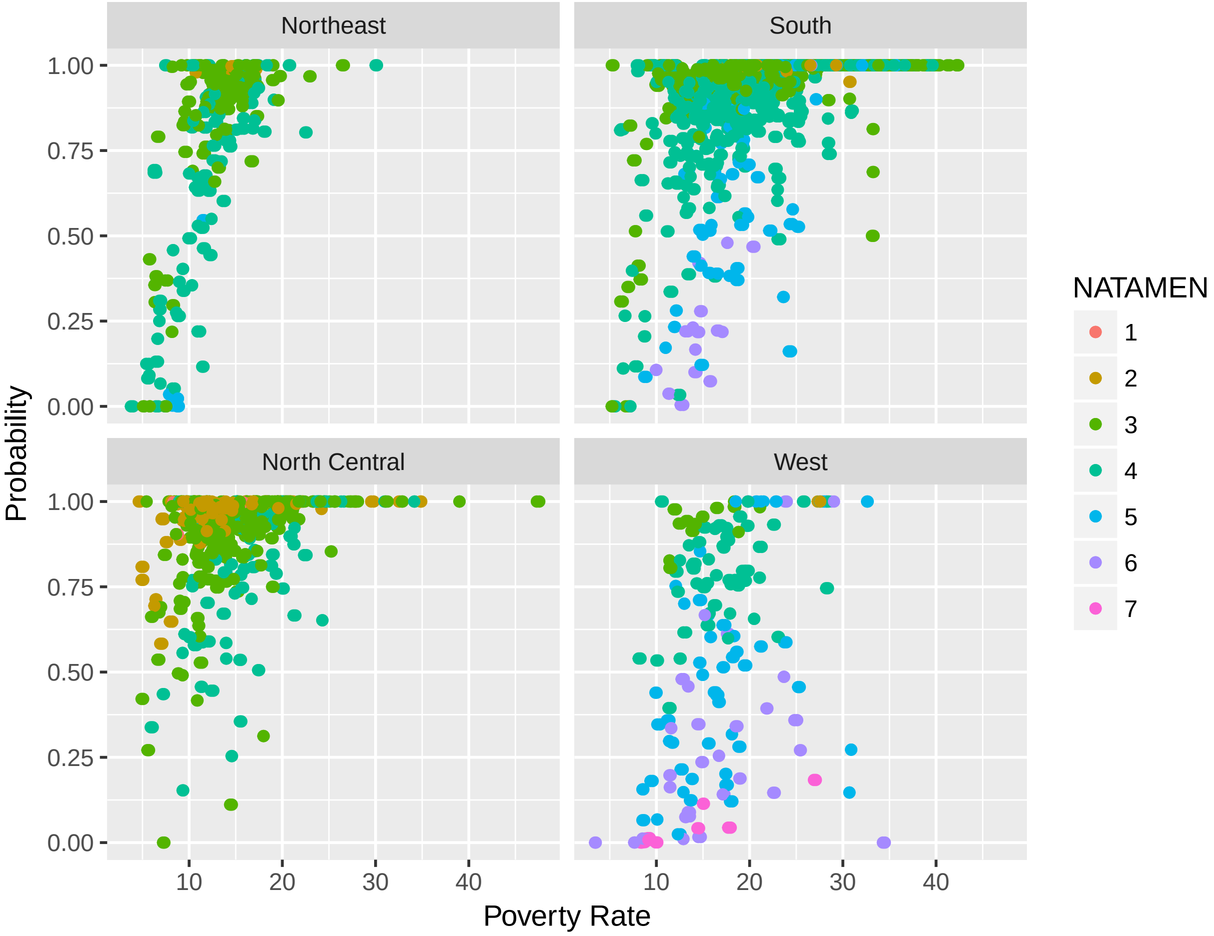
After examining the raw data constituting our model, we decided to formalize our predictive model to answer our larger question as stated earlier: are we able to reliably predict the probability of populations showing a greater than average obesity rate based just on the factors identified previously?



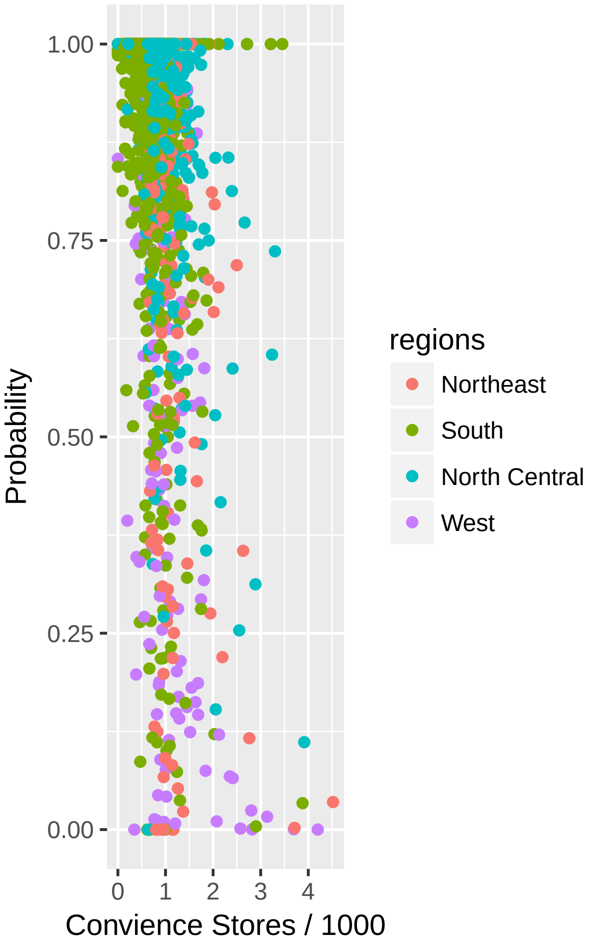
The above left plot, the first of our probabilistic predictions shows that the west generally has higher NATAMEN and lower probabilities of being above average for obesity, while the North Central region has the lowest NATAMEN and a high probability, the South region has a high probability and moderate NATAMEN, and the Northeast region also shows a moderate range of NATAMEN and generally high probabilities with some lower ones mixed in. In general the lowest probabilities are heavily clustered in the West and at high NATAMEN.

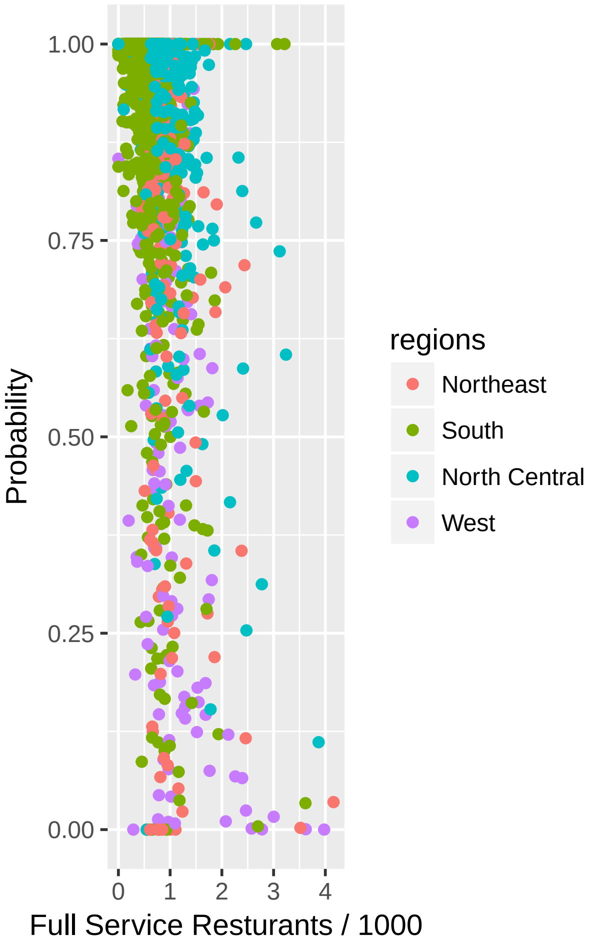


We see from this plot that the majority of counties show a moderate poverty rate, between 10 and 20% or so. Overall, it seems as though the Northeast as a region shows the lowest average rates of poverty, th South shows the highest, the West shows a generally low rate - but with a wide range, and the North Central region also shows a lower general rate with obvious, very high outliers. From this plot, it doesn't appear as though there is a strong relationship between either obesity, or NATAMEN and poverty rate. Though there are very few high poverty points with low probability, there isn't a completely consistent pattern of high poverty counties being assigned high obesity rates, or low NATAMEN values for that matter. This is heartening, as the naive, pessemistic assumption might be that our high poverty communities are all located in undesirable locations and the people living in them are destined to poor health outcomes.



This interesting plot shows that, while the West is quite complicated, the other three regions all show a striking relationship between increasing poverty rate and the probability of higher than average obesity rates. We also see that there are some horizontal "bands" of points colored by NATAMEN, with counties generally falling around a set probability dependent on their NATAMEN score, largely regardless of the poverty rate. This seems to mean that while poverty rate is important to the probabilities seen here, NATAMEN scores may play a more important role overall. This plot paints a decidely less optimistic picture than the previous one, as it appears that increasing poverty does correlate strongly with obesity in most cases, and the natural amenities of counties may be largely determining the fate of their inhhabitants.





These interesting plots show both that there are very real differences in probabilits between the regions, and that, as we identified earlier, there is an inverse relationship between the number of full service resturants and convenience stores. This relationship between stores/resturants and obesity is quite interesting and definitely counter to our a priori expectations. Unfortunately we do not have any clear ideas for why this is the case, as we generally expect that the prevelance of unhealthy food should correlate with an increase in obesity rates. One hypothesis could be that the number of stores/resturants isn't actually much related to the obesity rate, but the correlation between them seen is just due to the larger number of these businesses in the West region where rates are already lower for other reasons.

Conclusion and Limitations

As stated above, overall we are quite happy with our model and believe that we have addressed both of our standing questions. In regards to the first one, we found that NATAMEN, the number of full service resturants and convience stores were the best and strongest predictors of obesity rate. Using this information we were able to construct a model which quite reliably predicted the probability of communities showing a higher, or lower, than average obesity rate based just on those factors. We believe that our model uncovered some valuable and interesting insights which could be of potential use in the fight against obesity. Despite our successes, we were able to identify some very real limitations with our work.

Aside from the NA problem identified above, we were also unsure how the data were collected. Variables such as CONVSPTH07 and FSRPTH07 are likely to be accurate because this data is easy to collect by the government. However, it would be really nice to know how they collected the data for PCT_OBESE_ADULTS10 which seems like a hard variable to accurately measure.

As far as our model was concerned, we were able to fit a very good model using just three variables. However, there is a slight bump at the right end of our residual-fit plot. In an ideal world we could find another variable in our dataset that isn't diabetes which we could add in to fix this bump and get a perfectly straight line. We were not able to find one in our data set.

One variable that would have been interesting to study that wasn't available was a classification of FSRPTH07 into types of restaurants into food types (vegetarian, Thai, Steakhouse, Chinese, Italian, etc.) or a variable that quantifies the healthiness of that restaurant (a grade 1-5 based on the average calories contained in a meal or something like this). If we were able to classify the restaurants into groups and facet some of our plots we may have found some interesting information.

In general we believe that all of these shortcomings could have been addressed satisfyingly given more time, more data, and more variables to work with.