How Can We Predict Lyme Disease Using Two Predictor Variables: Economic Status and Education Level?

MGMT 6233: Introduction to Business AnalyticsSemester Project

Donaldson. MJ, Karaaslan. Ezgi

How Can We Predict Lyme Disease Using Two Predictor Variables: Economic Status and Education Level?

What is Lyme Diseaes?

Lyme disease, also known as Lyme Borreliois, is an infectious disease caused by the bacterium Borrelia burdoferi. It is transmitted to humans by the bite of infected blacklegged ticks (*aka.* deer tick). While found primarily in the Northeast Regions of the United States[7] cases have been reported in nearly all fifty states.

The most commonly recognized sign of infection is an expanding area of redness that can clear as it enlarges; leaving the classic "Bull's Eye" Rash known as Erythema migrans (EM) rash at the site of a tick bite; anywhere from 3-30 days after a bit has occurred. The rash is typically neither itchy nor painful. Other early symptoms may initially include fever, headache, and feeling tired. If untreated, symptoms may include arthritis, Lyme Carditis, Inflammation of the Brain and Spinal Cord, Bell's Palsey (loss of the ability to move one or both sides of the face), joint pain, neuropathy and short term memory loss[10].

Lyme disease is the most common disease spread by ticks in the Northern Hemisphere and estimated to affect 300,000 people a year in the United States. A previous vaccine is no longer available, however, research is ongoing to develop new vaccines. Antibiotics are the primary treatment for this disease.

Potential Business Applications using Lyme Disease Analytics

Pharmaceutical companies who sell those antibiotics would want to determine the right target to focus their marketing and selling activities on those places. They would also want to know the potential new markets to expand their businesses. The questions that they might seek for an answer could be:

- Where in the US is there the highest concentration of affected individuals?
- Do the counties most infected tend to be rural or city?
- Are there any relations between some of the socioeconomic parameters and the disease occurrences? And if so, can we predict the disease using some of those parameters?
- Do the places with high-income level have the highest concentration of affected individuals?

Based on this, we would like to conduct a data-science oriented study where these companies could utilize

the results for their business applications.

Identifying Predictors

We identified two predictors to use in our analysis: Economic status and education level. The average annual out of pocket expense for a person undergoing treatment for late stage Lyme Disease is ~$10,343. We picked economic status with the idea that companies marketing treatment options with higher cost would benefit from initially targeting an area of the country that is both highly endemic and also has a higher number of patients who will be able to afford the treatment. The other option would be to look into the mean income of endemic areas to try to find an association there.

We picked education level as the second predictor, because we think that less educated people might not be aware of the disease or how to treat them, therefore there is a potential market opportunity for the companies there. Additionally, it has been shown that those who spend greater amounts of time outdoors in their occupation, or recreationally are at much greater risk of coming into contact with a disease carrying tick. This is especially true for those who work out doors in woods or near large grassy fields [1,2] . The idea that education level often being associated with a person's profession led us to further explore this as a variable. More details on the reasons for choosing these variables are explained in the **Modeling** section.

As a result, our topic for this project is that *how can we predict Lyme disease using two predictor variables: economic status and education level?*

Data Understanding & Preparation

To conduct our analysis and to get the data we need, we gathered data from three different resources:

1. The Centers for Disease Control and Prevention: We were able to find data for Lyme disease prevalence by county for all fifty states for the years 2000-2014. This data was separated by state down to the county level.

2. Census Bureau – SAIPE data: We found data down to the county level for median household income for all fifty United States at the county level for the years 2009-2013.

3. United States Department of Agriculture Economic Research Service: We were able to find data for education level for all fifty United States at the county level. This data came aggregated for the years 2009-2013.

We consolidated these three sets of data into one document. As the next step, we selected what specific data we want to include, cleaned, and formatted the data. We were able to load this raw data to R as a .csv and we kept constructing the data there.

There was a challenge in the data in that the education data was aggregated for the years 2009-2013. The disease case and the income information, however, were available for the years 2009 to 2014. We were planning to use only one or two years for our analysis, but decided to aggregate both Lyme disease cases and income data for the years 2009-2013 as well, in order to remain consistent across all three datasets, and to assure that we were performing an "apples to apples" comparison.

We chose not to aggregate any additional factors, besides year, as we wished to retain as fine granularity as we could. We aggregate them by the same summary statistic: mean.

Finally, it was important to normalize case count to account for the possibility that disease count could be driven by population size. To do this, we adjusted the case variable to perform analysis for cases per 100,000 people.

A summary of our data:

Dependent Variable					
Label	**Variable**	**Mean**	**Median**	**Min.**	**Max.**
Cases2009, Cases2010, etc.	Lyme disease case counts for each county for the years 2009 to 2013.				
AggregateCasesPerOneHundredThousand	Created for aggregation and normalization	9.7	0	0	478.8

Independent Variable					
Label	**Variable**	**Mean**	**Median**	**Min.**	**Max.**
Less than a high school diploma, 2009-2013	Educational attainment for adults age 25 and older for the U.S. states and counties	9,191	2,831	4	1,510,000

High school diploma only, 2009-2013		18,480	6,399	18	1,324,000
Some college or associate's degree, 2009-2013		19,100	5,019	12	1,706,000
Bachelor's degree or higher, 2009-2013		18,960	2,796	6	1,916,000
POPESTIMATE2009 POPESTIMATE2010, etc.	County population information for each year				
meanAggregatedPopulation	Created for aggregation of the years 2009 – 2013	99,140	25,750	79	9,906,000
Median_Household_Income_2009 Median_Household_Income_2010, etc.	Median Household income for the years 2009 – 2013				
AggregatedIncome	Created for aggregation of the years 2009 – 2013	44,160	42,260	21,490	117,900

Fig. 1

Independent and Dependent Variable

We wanted to predict whether education and/or economic status affect the prevalence of Lyme disease. For this reason, we used education and the income as the independent variables and cases of Lyme disease as the dependent variable.

We conducted an exploratory analysis to make a comparison between states. The detailed code and the results are in **Appendix I**.

Arkansas and Hawaii have zero cases per 100,000 people. Minnesota, Pennsylvania, and Wisconsin are the top three states that have the highest cases. We used the "order" function to sort the data and the "aggregate" function to have the data at state level.

Modeling

We had three categories of quantitative data that we wished to evaluate. The goal was to determine whether education, median household income, or a combination of the two could be used to identify higher risk areas.

It is important to note that our theory was not that less education, or lower income was causal for Lyme Disease. Instead, we were attempted to identify relationships between these variables

i.e. we would like to know if people with high school diploma are more likely to get sick or counties with higher income have more disease cases. More details on these variables are explained below.

For these reasons, we chose linear regression analysis[5] for our modeling technique.

In cases where it made sense to do so, we plotted our regression along with a scatterplot to create a visual representation of how the data fell.

Why education could be a predictor of Lyme disease?

The question we want to answer in this regression is "Is education level a good predictor of populations more likely to be infected by Borrelia burgdorferi?" There are several reasons to think this may be a factor. As mentioned briefly above, the current CDC model focuses on tick population studies to identify potentially endemic areas. While living in an endemic area has been shown to increase a person's risk of infection, additional risk factors have also been shown.

In one study (Smith et al.), it was shown that age, living in a single family home, homes with yards or attached land especially if they are within 100 feet of woodland, and gardening more than four hours a week all contribute to an increased probability of becoming infected.

In 2002, OSHA published a Hazardous Information Bulletin (HIB) to warn people in endemic areas of increased risk associated with certain jobs. The examples given were farming, land surveying, railroad work, landscaping, forestry, park/wildlife management, brush clearing, oil field work, utility line work, and construction work.
Many of the aforementioned jobs offer entry-level positions that do not require education beyond college. In fact, in rural areas, it is common for young children to help gather crops on family farms. This led to the idea that a person's level of education may serve as a predictor as they are more likely to work in a higher risk line of work.

Why income could be a predictor of Lyme disease?

A study (Linard et al.) conducted in Belgium in 2007 found that both environmental and socio-economic factors could be used to explain spatial variation in disease risk.

The working theory of this study was that forest workers, farmers, foresters and loggers had a greater exposure to risk factors and that these professions tend to correspond with lower mean incomes.

For this same reason, we chose to evaluate the mean of the median income by county for data aggregated

for the years 2009-2013 to determine whether or not it may serve as a predictor.

Regression Details:

- As mentioned before, we converted the number of cases to "cases per 100,000 people". We believe that this would give us a larger coefficient. It was also important in properly scaling the data.

- Education data was pre-aggregated for the years 2009-2013 and divided into four categories: Less Than a High School Diploma, High School Diploma Only, Some College or Associate's Degree, Bachelor's Degree or Higher. In this case to properly scale the data, raw education counts were divided by the mean population by county for 2009-2013 to gain a percentage by county of the four separate education levels.

- We used multi-level regression model to include states as a grouping level.

  Multi-level models involve predictors from multi-levels and their interactions. They must account for associations among observations within levels to make efficient and valid inferences.

  Regular regression ignores the average variation between entities and individual regression may face sample problems and lack of generalization.

- Income data is a dollar value and values are often very large numbers, therefore, we decided to use log transformation. This enabled us to interpret the regression coefficient as a percentage change instead of a "one unit" change.

  We ran income model and compare it with another model that uses log (income) instead. Log-transformation, however, didn't increase the model fit, so we will use the untransformed regression (see **Appendix II**).

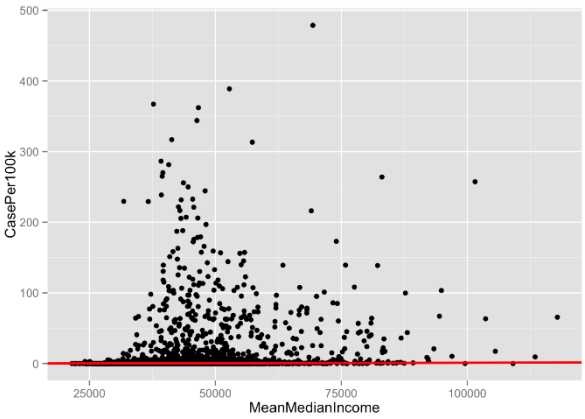
Income Regression for all 50 States Results and Evaluation

The first regression we ran was for the income variable (see **Appendix III**).

The R^2 value was 0.4521; meaning that only 45.21 of the variance within the data could be explained by the regression. We noted that standard error was very high (3.622e+00).

The p-value for the variable was <2.2e-16 which would be good. However, the t-value was very low (-0.050) indicating that the fit was not very precise, nor did it explain very much about the data. These results combined with the high standard error showed us that income as a predictor variable is insignificant.

The reason behind this might be that larger counties have more cases and we would expect a county with more people to have higher income levels. However, we are unable to confirm a normal distribution of cases. So, it is possible that another explanation exist. We will discuss the distribution of raw case counts later.



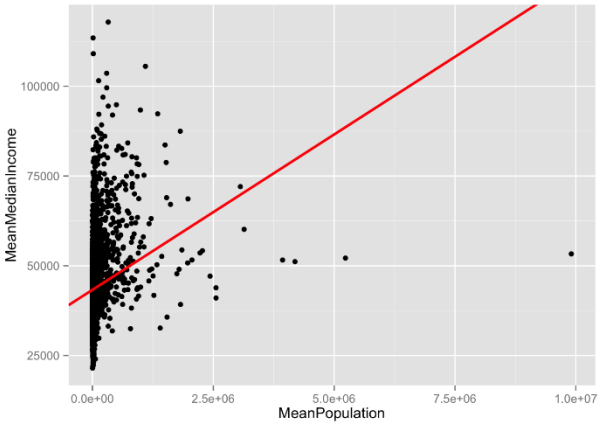
Income ~ Case Count Regression Using data for all 50 United states- It can be seen here that the slope of the regression is near zero. Indicating no significant relationship between income and Lyme Diseae Case Counts.

As a secondary test, we wanted to investigate the relationship between county population and income. (**Appendix IV**).

The R^2 value of this regression was 0.06114 showing a significant decrease in the amount of variance explained. Here again, the standard error was high (2.012e+02).

The p value was <2.2e-16 indicating that we could reject a null hypothesis. The t-value was much higher at 215.2 The relationship of this regression seemed to be significant and positive, meaning that disease cases are only a function of county size and not of income. As the p-value is very low, we can say that our measure of income is only a proxy for county size. However, only 6.1% of the variance is represented. So, it is likely that there is further explanation here as well, but

that a relationship to Lyme Disease Cases is not present.



Population ~ Income Regression Using data for all 50 United States – here a direct relationship can be seen between population and income.

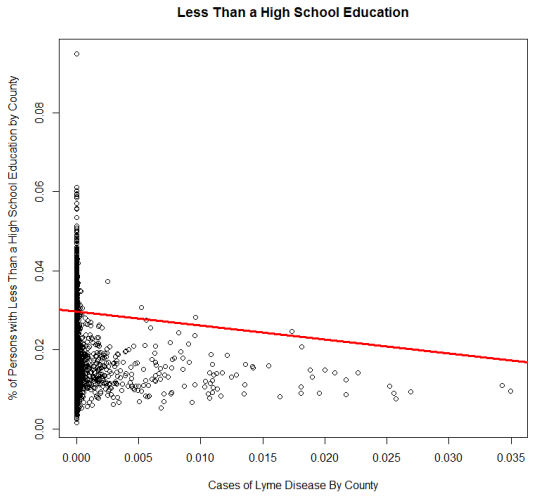
Our theory was that income level would be a significant predictor for the Lyme disease cases for the counties. Certain professional or leisure activities that are dominant in certain counties could be reflected in the socio-professional level.

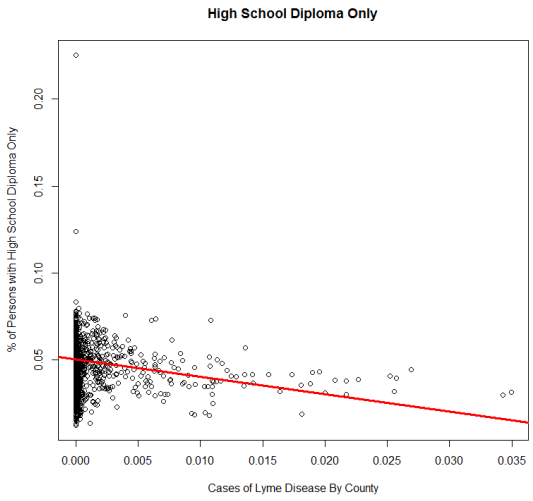
Our results showed us that this predictor is not a significant one and the model wouldn't be very useful.

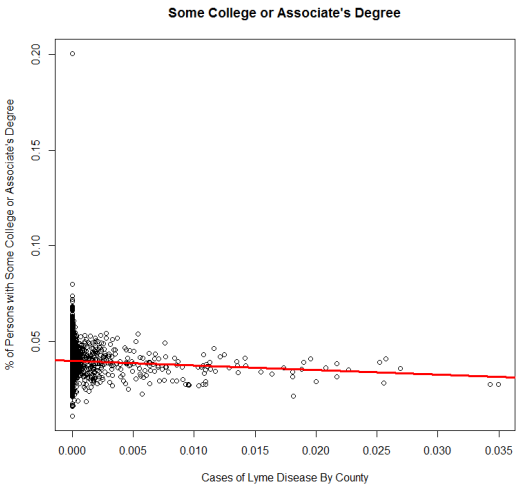
How data was sorted for Education:

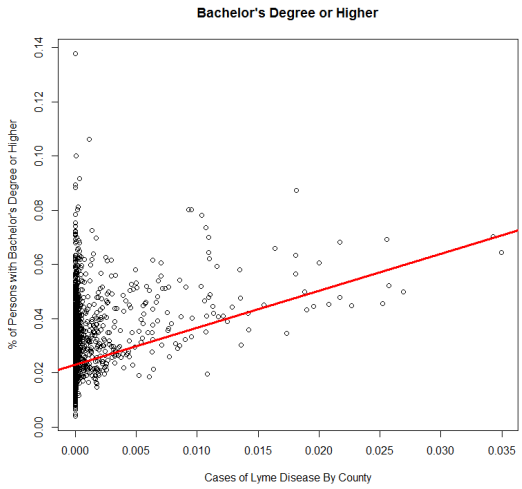
As mentioned above, Education data was pre-aggregated for the years 2009-2013 and divided into the categories Less Than a High School Diploma, High School Diploma Only, Some College or Associate's Degree, Bachelor's Degree or Higher.

These categories were first renamed so that they could be easily used as variables. Once the data was cleaned and properly scaled, we performed the linear regression. Our results were as follows:









Interpretation of results:

R^2 values for this data were much higher for Less That a High School Education and Some College or Associate's Degree.. These variables represented 40% and 30% of the data respectively with standard errors of 0.0008782 and 0.0008130.

While less significant, High School Diploma only accounted for 28% of the data with a standard error of 0.0011830. Bachelor's degree or higher explained the least amount of variance (25%) with a standard error of 0.0012667.

P-values for all four categories were low (< 0.001) with t-values in the range 18.213 - 49.018. We are testing for the null hypothesis that the education variables have no effect. A p-value close to zero allows us to reject the null hypothesis[12]. In all four categories, we were reject the null hypothesis, this allows us to state that the education variables 'affect' Lyme Disease. Again, we are using the term 'affect' lightly as we are not attempting to identify new causal factors, but instead we are looking for relationships that may exist.

T-values are a measure of the standard error to coefficient ratio. The larger ratios shown in our results indicate that not only are the education variables likely significant, but that that the results are precise enough to be considered valuable

The first three categories representing populations with less than a Bachelor's degree all show a negative slope, demonstrating an inverse relationship. As the number of Lyme Disease Cases rises for these populations, it is less likely that a person will fall into one of these three categories.

For the category representing persons with a Bachelor's Degree or Higher, the slope was positive, demonstrating a relationship that is directly proportional. As the number of cases rise, it can be expected that a greater

number of people in this category will be represented.

These were interesting results, however, we did note that there were rather significant outliers that seemed to be weighting the regression in one direction or the other. We wondered how this data would look if we were to remove this element. Which, we will discuss this later.

Combining Education and Income for All 50 United States:

The same study (Linard et al), mentioned above in the "Why education could be a predictor of Lyme disease?" section, conducted multivariate statistical analysis. We were inspired by this approach and decided to run a multiple linear regression using the same income data and education levels.

When this regression was performed, the R^2 value decreased significantly across the board with all categories explaining between 8% to 10% of the variance. As would be expected, standard error also rose significantly.

P-values for all four education variables were < 0.05, showing that education alone is a better predictor than education paired with median income.

T-values for the four education variables showed a much wider range

(-3.522 - +5.797). Again, larger rations allow us to make positive statements about the significance and precision of the data. In this case, the four variables were associated with small t-values. The variable with the highest ratio was High School Diploma Only, however, this variable could only be associated with 10% of the data.

The values seen in this particular regression very likely goes back to the values for Median Income seen above. Further confirming that there is no significant relationship between income and Lyme Disease Risk.
We decided to reject this model.

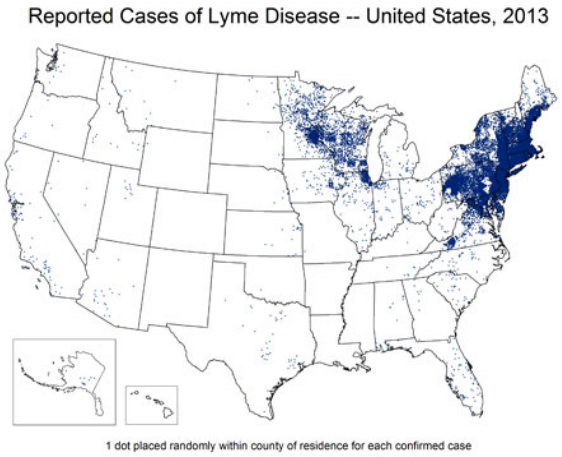


Fig. 1
The bacterium *Borrelia burgdorferi* is reliant of the mutation of ospC, which occurs in the midgut of blacklegged ticks (deer tick)[8], to become infectious.

Populations of the blacklegged tick are higher in the Northeast where a direct correlation between cases of *B. burgdorferi* to deer tick populations can be seen[7].

Regression of Education in Endemic Regions:

As mentioned above, we noted significant outliers in our data set. All regressions were performed using State Name as a factor. The result of which gave a coefficient for each state. It was noted that there a high degree of variance existed between the t-values from state to state. This was not overly surprising as the distribution of Lyme Disease is much greater for Northeast Regions of the United States (fig1).

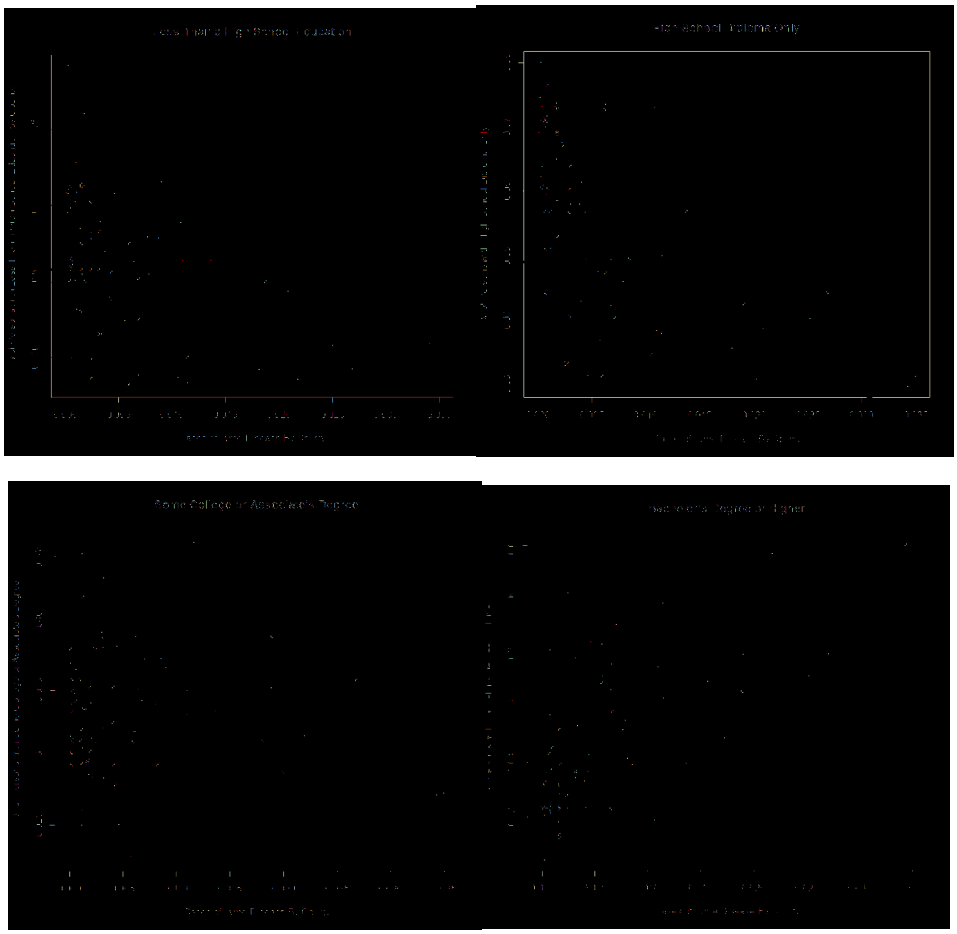
With this in mind, we decided to perform linear regressions using data for three of the most endemic states for Lyme Disease; MA, PA, and CT. Despite the raw counts mentioned about, we chose these specifically as a representation of states densely populated with Lyme Disease patients.

Interpretation of Results:

As expected, the slope of the regressions did not change. Categories represents those with less than a Bachelor's degree remained negative, while the category representing individuals with a Bachelor's Degree or Higher remained positive.

The R^2 values for both Less than a High School Diploma and Some College or Associate's Degree decreased significantly, while those with a High School Diploma Only and those with a Bachelor's Degree or Higher both more than double.

This was rather surprising. Our thought as to why these values could have changed so significantly in their individual categories is that we are likely seeing a reflection of higher high school and college graduation rates in these areas. R^2 values for all for education variables ranged between 0.12 and 0.60.



For ease, these values can be see in Fig 2. below.

Fig 2 R-squared values for endemic vs non-endemic states

	Less than High School Diploma	High School Diploma Only	Some College or Associate's Degree	Bachelor's Degree or Higher
All 50 United States	0.40	0.28	0.38	0.25

MA, PA, and CT only	0.15	0.60	0.12	0.56

A possible explanation for this is that there are much higher rate of both high school and undergraduate completion in these three states.

Standard error for all four categories was very low; ranging from 0.0016771 to 0.003724.

P-values for this regression were < 0.001. Again, suggesting that education may serve as a valuable predictor of Lyme Disease. T-values were reduced with a range of 9.309-22.178. Suggesting that the model using all 50 states is more precise.

The results of this regression showed that the previous fit, while covering a wider range of case counts, was not skewed by these outliers. We came to this conclusion by the fact that the slope of our regression lines were not radically altered when run with a smaller data set representing a higher percentage of cases.

Education and Median Income for Endemic States:

To follow through on our analysis, we evaluated both education and median income as predictors for Lyme Disease cases in the same three endemic states as above for our last regression.

Again we saw a similar trend between non-endemic and endemic regions. R^2 values remained between 8% and 10%. However, standard error was not affected as it had been in the previous regression using all 50 states. The range for standard error was 1.034e-01 - 0.10452

The interpretation of these results is that the combination of education and median income only represents 8%-10% of the data, but that this regression is more reliable as it has a lower standard error.

P-values for all four education/income variables remained at < 0.05. T-values were more consistent, displaying a tighter range, however, these values were still very small, ranging from 2.242 - 2.534.

The results above led us to again reject the combination of education and median income as a valuable predictor. This regression explains a small amount of the data and cannot be said to be precise.

What we might have chosen instead:

Linear regressions assume variables are derived from a normal distribution[1]. Disease data like that of

our Lyme Disease cases often follows a Poisson distribution.

The Belgium study (Linard et al.) found that their own data showed much larger variances than the means of human infections.

This points to the possibility that our results are displaying over dispersion. In other words, we are seeing a higher degree of variability in the data than would be expected with a normal distribution. The solution in the study mentioned above was to add an overdispersion parameter.

The DCluster package for R can be used to perform Likelihood Ratio tests and Dean's tests, which may be appropriate for applications such as ours[4]. However, this is outside the scope of this project.

Conclusion:

Results for Median Household Income were not promising as potential predictors of Lyme Disease. The significant values for this particular variable were low enough that it is unlikely to be valuable even with a more complex approach.
The results for education as a predictor were much more promising. Further analysis using more complex, multivariate analysis could reveal more about the true value of this variable. It would be important to account for the possibility of a Poisson distribution and to consider using population as an offset variable.

Deployment

The concept of deployment in predictive data mining refers to the application of a model for prediction to new data. Building a model is generally not the end of the project.

Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process.

In our case, customers are the pharmaceutical companies who would be able to use the model for their business decisions.

The applications can vary, however, the main utilization can be for predicting the disease cases to be able to forecast the demand. This way, they can plan their marketing and sales activities more efficiently. In addition, other information gained from the analysis can be utilized as well:

Scenario 1) Dr. Richard Horowitz is a big fan of data analytics. He has hired data analysts to perform analytics on over 10,000 patient records spanning more than a decade to determine commonalities between his patients. He has compiled this information and his own clinical experiance as one of the country's top Lyme Disease Specialist to write the book Why Can't I get Better? .

The Horowitz team would now like to extend their use of data analytics to identify target markets for Dr. Horowitz's books to the community's throughout the United States Counties most in need.

Before these, though, the model should be tested and validated. It can be done, for example, by using actual numbers of year 2014 and comparing them with the regression results. Additionally, the users should avoid overfitting and learn how to recognize it by using different analytical tools.

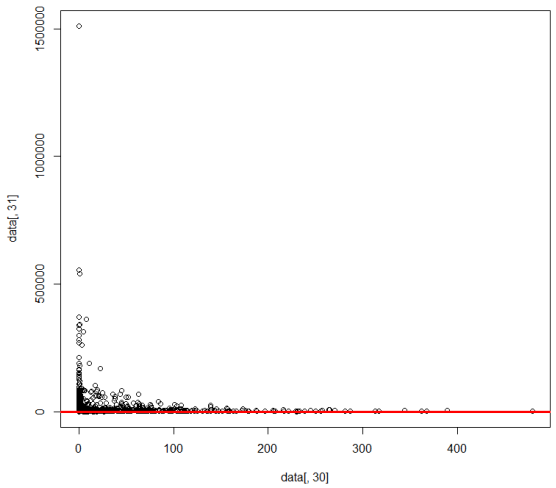
Scenario 2) Tick Twister is a useful tool in the prevention of Lyme Disease. This specialized product is used to cleanly remove ticks, including the tick head, in the event that someone finds a tick attached after a hike, or working in the garden.

The makers of Tick Twister have primarily focused on Lyme Disease Conferences, but have found diminishing returns in the last few year. They believe this is largely because the same patients and practitioners come to the same conferences year after year. They would like to use data analytics to identify new target markets in the United States.

Scenario 3) The Centers for Disease Control and Prevention has increased its estimation of new Lyme Disease each year 10 fold from 30,000 each year to a newly estimated 300,000 [13,14].

The new estimate has put pressure on the CDC to find new and novel approaches to identifying endemic areas.
If further analysis were to show education variables to be truly significant, these could be paired with current data to enhance the prediction of emerging endemic areas and to better solidify Lyme Disease estimates in the fourteen states that have already been identified.

In this particular data set, raw case counts were plotted against mean education values.



Looking forward

The dataset created for this project proved to be useful in many ways. It allowed us to gain a real world working experience with data that needs to be shaped prior to analysis. Working with files that could easily be converted to a .csv allowed us to focus on this shaping and the regressions themselves.
For future projects, in would be helpful to first research the type of model the a particular data set tends

to follow. We were fortunate that a linear regression could be used to explain at least some of the data, however, had we known that case count data tends to follow a Poisson model instead of a normal distribution, we would have either started with a different approach to the data, or chosen a different data set that followed a normal distribution. Of course, this is not always an option in real world scenarios. So, there was value in this lesson in that it is important to have and understanding of the data and the model it likely follows when designing a project.

What we Learned

This project really opened our eyes to the importance of properly scaling data prior to performing a regression. This was immediately apparent the first time we plotted a regression without the proper scale. In the figure shown below (fig. 3), it is possible to see the nearly flat regression line running through the data.

In this particular data set, raw case counts were plotted against mean education values.

One of the main challenges that we ran into One of the main challenges we ran into was how exactly to scale data. We first attempted to apply a log function to education data, however, we saw no improvement by doing so.

With the goal of getting values across the board near the range of 0-1, it was decided that we should evaluate cases per 100,000 and education and income as a percentage as explained above.

We also determined that it was important to take the mean values of aggregated data in order to keep values consistent across case count, income and education data.

A suggestion would be to pair significant variables such as education with tick counts.

References:

[1]Smith, G., E. P. Wileyto, R. B. Hopkins, B. R. Cherry, and J. P. Maher. "Risk Factors for Lyme Disease in Chester County, Pennsylvania." *Public Health Reports (Washington, D.C. : 1974)* 116 Suppl 1 (2001): 146. Print.

[2]Occupational Safety and Health Administration, United States. "Lyme Disease Facts." Washington, D.C.]: Washington, D.C. : U.S. Dept. of Labor, Occupational Safety and Health Administration, 2002. N. pag. Print.

[3]Venesky, Tom. "Game Commission to Add Lyme Disease Education to Hunting Course." *Times Leader (Wilkes-Barre, PA)*. N.p.: n.p., 2014. N. pag. Print.

[4]Linard, Catherine, Penelope Lamarque, Paul Heyman, Genevieve Ducoffre, Victor Luyasu, Katrien Tersago, Sophie O. Vanwambeke, and Eric F. Lambin. "Determinants of the Geographic Distribution of Puumala

Virus and Lyme Borreliosis Infections in Belgium." *International Journal of Health Geographics* 6.15 (2007): 15. Print.

[5] Linear Regression. (n.d.). Retrieved December 7, 2015, from http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

[6](2014, August 27). Retrieved November 17, 2015, from http://www.cdc.gov/lyme/stats/maps/interactiveMaps.html

[7]Transmission. (2015, March 4). Retrieved November 14, 2015, from http://www.cdc.gov/lyme/transmission/index.html

[8]Borrelia burgdorferi OspC protein required exclusively in a crucial early stage of mammalian infection. (n.d.). Retrieved November 14, 2015, from http://www.ncbi.nlm.nih.gov/pubmed/16714588

[9]DAVIDIAN, M. (n.d.). Generalized linear models for nonnormal response. Retrieved December 14, 2015, from http://www.stat.ncsu.edu/people/davidian/courses/st732/notes/chap11.pdf

[10]Signs and Symptoms of Untreated Lyme Disease. (2015, August 17). Retrieved November 14, 2015, from http://www.cdc.gov/lyme/signs_symptoms/

[11]LYMEPOLICYWONK: Annual Lyme costs now top $3.1 billion-It's time to wake up! - LymeDisease.org. (2013, November 4). Retrieved December 12, 2015, from https://www.lymedisease.org/lymepolicywonk-annual-lyme-costs-now-top-3-1-billion-its-time-to-wake-up-2/

[12]How to Interpret Regression Analysis Results: P-values and Coefficients | Minitab. (n.d.). Retrieved December 12, 2015, from http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-regression-analysis-results-p-values-and-coefficients

[13]How many people get Lyme disease? (2015, September 30). Retrieved December 14, 2015, from http://www.cdc.gov/lyme/stats/humancases.html

Appendices:

APPENDIX I

```
setwd("/users/ezgikaraaslan/Desktop/Data Analytics/Project")
rawdata <- read.csv("Raw_Data_3.csv", header=TRUE)

rawdata$Population2009To2013 <-
rawdata$POPESTIMATE2009+rawdata$POPESTIMATE2010+rawdata$POPESTIMATE2011+rawdata$POPESTIMATE2012+rawdata$POPESTIMATE2013
rawdata$MeanPopulation <- rawdata$Population2009To2013/5
rawdata$Cases2009To2013 <-
rawdata$Cases2009+rawdata$Cases2010+rawdata$Cases2011+rawdata$Cases2012+rawdata$Cases2013
rawdata$MeanCase <- rawdata$Cases2009To2013 / 5
rawdata$CasePer100k <- (rawdata$MeanCase*100000) / rawdata$MeanPopulation
rawdata$MedianIncome2009To2013 <-
rawdata$Median_Household_Income_2009+rawdata$Median_Household_Income_2010+rawdata$Median_Household_Income_2011+rawdata$Median_Household_Income_2012+rawdata$Median_Household_Income_2013
rawdata$MeanMedianIncome <- rawdata$MedianIncome2009To2013/5
summary(rawdata$CasePer100k)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   9.666   1.809 478.800

summary(rawdata$Less.than.a.high.school.diploma..2009.2013)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       4    1130    2831    9191    6352 1510000

summary(rawdata$High.school.diploma.only..2009.2013)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      18    2800    6399   18480   15170 1324000

summary(rawdata$Some.college.or.associate.s.degree..2009.2013)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      12    2226    5019   19100   13690 1706000

summary(rawdata$Bachelor.s.degree.or.higher..2009.2013)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       6    1094    2796   18960    9078 1916000

summary(rawdata$MeanPopulation)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      79   11020   25750   99140   66920 9906000

summary(rawdata$MeanMedianIncome)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   21490   36780   42260   44160   49070  117900       2
```

```
ordercases<-order(rawdata$CasePer100k)
dat<-data.frame(STNAME=rawdata$STNAME[ordercases], Index=rawdata$Index[ordercases],
Cases=rawdata$CasePer100k[ordercases])
agg <- aggregate( Cases ~ STNAME, dat, sum)
ordercases2<-order(agg$Cases)
dat2<-data.frame(STNAME=agg$STNAME[ordercases2], Cases=agg$Cases[ordercases2])
dat2
```

```
##              STNAME       Cases
## 1          Arkansas    0.0000000
## 2            Hawaii    0.0000000
## 3          Colorado    0.1288221
## 4       Mississippi    2.0407299
## 5        New Mexico    3.2850351
## 6  District of Columbia    4.4454724
## 7          Oklahoma    4.4569010
## 8              Utah    4.6093776
## 9           Arizona    4.9023900
## 10        Louisiana    5.5909813
## 11          Wyoming    7.6235056
## 12         Missouri    8.2155559
## 13     South Dakota   10.3528655
## 14       Washington   12.7227946
## 15         Kentucky   17.2839696
## 16          Alabama   19.4052980
## 17            Idaho   22.8635139
## 18           Alaska   25.5384612
## 19          Montana   31.6750761
## 20   South Carolina   33.2000980
## 21         Nebraska   35.5888782
## 22           Nevada   37.5967409
## 23          Georgia   38.1587776
## 24          Florida   41.7731367
## 25           Oregon   45.4519401
## 26             Ohio   45.6710682
## 27           Kansas   49.2953263
## 28       California   52.3231154
## 29        Tennessee   58.8947834
## 30     North Dakota   93.8333833
## 31          Indiana  121.4750752
## 32   North Carolina  151.8665282
## 33            Texas  178.3831597
## 34     Rhode Island  178.7614610
## 35         Illinois  202.6745218
## 36         Delaware  223.5741835
## 37    West Virginia  312.8497893
## 38         Michigan  316.1080468
## 39             Iowa  430.8876945
```

```
## 40        Connecticut  684.1250831
## 41      New Hampshire  829.0754194
## 42           Maryland 1094.5906673
## 43            Vermont 1221.8982457
## 44              Maine 1256.8988561
## 45         New Jersey 1509.0882906
## 46      Massachusetts 1581.0167181
## 47           Virginia 1619.3473483
## 48           New York 2879.9410283
## 49          Minnesota 3218.6824431
## 50       Pennsylvania 4567.2998276
## 51          Wisconsin 7083.8168055
```

APPENDIX II

```
library(texreg)
regr1 <- lm(CasePer100k ~ MeanMedianIncome, data=rawdata)
regr2 <- lm(CasePer100k ~ log(MeanMedianIncome), data=rawdata)
screenreg( list(regr1, regr2), omit.coef="factor", custom.model.names=c("Income Regression", "log(Income)
Regression"), include.rsquared=FALSE, include.rmse=FALSE)
```

```
##
## ===============================================================
##                    Income Regression  log(Income) Regression
## ---------------------------------------------------------------
## (Intercept)           -14.63 ***       -270.55 ***
##                        (2.48)           (27.85)
## MeanMedianIncome         0.00 ***
##                        (0.00)
## log(MeanMedianIncome)                    26.27 ***
##                                         (2.61)
## ---------------------------------------------------------------
## Adj. R^2                0.03113          0.03095
## Num. obs.            3141             3141
## ===============================================================
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

APPENDIX III

```
regr3 <- lm(CasePer100k ~ MeanMedianIncome + factor(STNAME), data=rawdata)
summary(regr3)
```

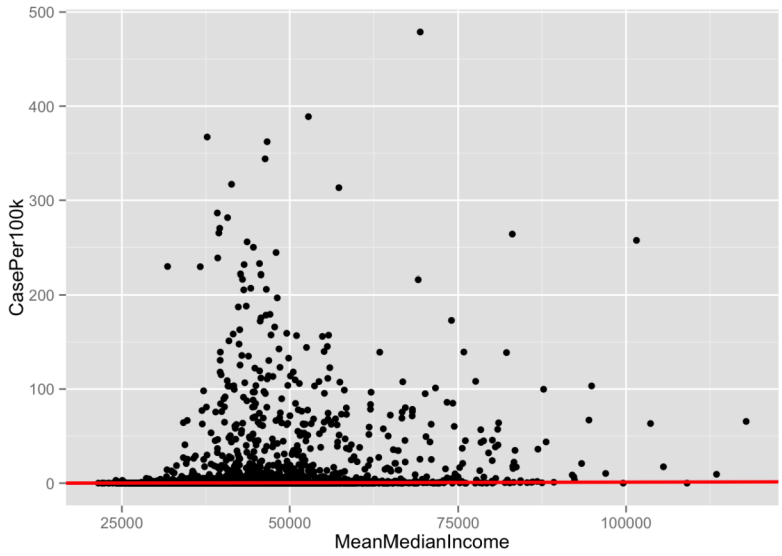
```
##
## Call:
## lm(formula = CasePer100k ~ MeanMedianIncome + factor(STNAME),
##     data = rawdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -100.46   -0.82   -0.23    0.05  365.75
```

```
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -1.818e-01  3.622e+00  -0.050  0.95996
## MeanMedianIncome                1.283e-05  4.902e-05   0.262  0.79351
## factor(STNAME)Alaska            3.511e-01  5.790e+00   0.061  0.95165
## factor(STNAME)Arizona          -2.039e-02  7.349e+00  -0.003  0.99779
## factor(STNAME)Arkansas         -2.716e-01  4.324e+00  -0.063  0.94992
## factor(STNAME)California        4.094e-01  4.677e+00   0.088  0.93025
## factor(STNAME)Colorado         -4.497e-01  4.538e+00  -0.099  0.92107
## factor(STNAME)Connecticut       8.484e+01  9.734e+00   8.715  < 2e-16
## factor(STNAME)Delaware          7.401e+01  1.520e+01   4.868 1.18e-06
## factor(STNAME)District of Columbia  3.823e+00  2.594e+01   0.147  0.88284
## factor(STNAME)Florida           2.673e-01  4.451e+00   0.060  0.95212
## factor(STNAME)Georgia          -8.406e-02  3.748e+00  -0.022  0.98211
## factor(STNAME)Hawaii           -5.585e-01  1.328e+01  -0.042  0.96645
## factor(STNAME)Idaho             1.511e-01  4.999e+00   0.030  0.97589
## factor(STNAME)Illinois          1.554e+00  4.081e+00   0.381  0.70329
## factor(STNAME)Indiana           9.035e-01  4.159e+00   0.217  0.82803
## factor(STNAME)Iowa              3.919e+00  4.106e+00   0.954  0.33994
## factor(STNAME)Kansas            7.764e-02  4.040e+00   0.019  0.98467
## factor(STNAME)Kentucky         -1.600e-01  3.923e+00  -0.041  0.96746
## factor(STNAME)Louisiana        -2.389e-01  4.497e+00  -0.053  0.95764
## factor(STNAME)Maine             7.818e+01  7.163e+00  10.915  < 2e-16
## factor(STNAME)Maryland          4.496e+01  6.270e+00   7.170 9.32e-13
## factor(STNAME)Massachusetts     1.123e+02  7.648e+00  14.689  < 2e-16
## factor(STNAME)Michigan          3.448e+00  4.232e+00   0.815  0.41528
## factor(STNAME)Minnesota         3.653e+01  4.234e+00   8.628  < 2e-16
## factor(STNAME)Mississippi      -2.263e-01  4.238e+00  -0.053  0.95741
## factor(STNAME)Missouri         -2.506e-01  3.955e+00  -0.063  0.94948
## factor(STNAME)Montana           2.294e-01  4.660e+00   0.049  0.96074
## factor(STNAME)Nebraska         -8.707e-03  4.139e+00  -0.002  0.99832
## factor(STNAME)Nevada            1.727e+00  7.024e+00   0.246  0.80575
## factor(STNAME)New Hampshire     8.237e+01  8.768e+00   9.395  < 2e-16
## factor(STNAME)New Jersey        7.116e+01  6.622e+00  10.746  < 2e-16
## factor(STNAME)New Mexico       -2.133e-01  5.470e+00  -0.039  0.96889
## factor(STNAME)New York          4.597e+01  4.588e+00  10.021  < 2e-16
## factor(STNAME)North Carolina    1.181e+00  4.064e+00   0.291  0.77137
## factor(STNAME)North Dakota      1.323e+00  4.766e+00   0.278  0.78139
## factor(STNAME)Ohio              1.144e-01  4.193e+00   0.027  0.97824
## factor(STNAME)Oklahoma         -2.788e-01  4.300e+00  -0.065  0.94832
## factor(STNAME)Oregon            8.831e-01  5.326e+00   0.166  0.86830
## factor(STNAME)Pennsylvania      6.775e+01  4.471e+00  15.153  < 2e-16
## factor(STNAME)Rhode Island      3.515e+01  1.198e+01   2.934  0.00338
## factor(STNAME)South Carolina    4.122e-01  4.925e+00   0.084  0.93330
## factor(STNAME)South Dakota     -2.210e-01  4.473e+00  -0.049  0.96059
## factor(STNAME)Tennessee         3.070e-01  4.104e+00   0.075  0.94038
## factor(STNAME)Texas             3.330e-01  3.545e+00   0.094  0.92518
## factor(STNAME)Utah             -3.195e-01  5.762e+00  -0.055  0.95579
## factor(STNAME)Vermont           8.684e+01  7.579e+00  11.457  < 2e-16
## factor(STNAME)Virginia          1.161e+01  3.910e+00   2.970  0.00300
## factor(STNAME)Washington       -1.042e-01  5.208e+00  -0.020  0.98403
## factor(STNAME)West Virginia     5.393e+00  4.679e+00   1.152  0.24921
## factor(STNAME)Wisconsin         9.795e+01  4.399e+00  22.265  < 2e-16
## factor(STNAME)Wyoming          -1.842e-01  6.275e+00  -0.029  0.97659
```

```
##
## (Intercept)
## MeanMedianIncome
## factor(STNAME)Alaska
## factor(STNAME)Arizona
## factor(STNAME)Arkansas
## factor(STNAME)California
## factor(STNAME)Colorado
## factor(STNAME)Connecticut         ***
## factor(STNAME)Delaware           ***
## factor(STNAME)District of Columbia
## factor(STNAME)Florida
## factor(STNAME)Georgia
## factor(STNAME)Hawaii
## factor(STNAME)Idaho
## factor(STNAME)Illinois
## factor(STNAME)Indiana
## factor(STNAME)Iowa
## factor(STNAME)Kansas
## factor(STNAME)Kentucky
## factor(STNAME)Louisiana
## factor(STNAME)Maine              ***
## factor(STNAME)Maryland           ***
## factor(STNAME)Massachusetts       ***
## factor(STNAME)Michigan
## factor(STNAME)Minnesota          ***
## factor(STNAME)Mississippi
## factor(STNAME)Missouri
## factor(STNAME)Montana
## factor(STNAME)Nebraska
## factor(STNAME)Nevada
## factor(STNAME)New Hampshire       ***
## factor(STNAME)New Jersey          ***
## factor(STNAME)New Mexico
## factor(STNAME)New York           ***
## factor(STNAME)North Carolina
## factor(STNAME)North Dakota
## factor(STNAME)Ohio
## factor(STNAME)Oklahoma
## factor(STNAME)Oregon
## factor(STNAME)Pennsylvania        ***
## factor(STNAME)Rhode Island        **
## factor(STNAME)South Carolina
## factor(STNAME)South Dakota
## factor(STNAME)Tennessee
## factor(STNAME)Texas
## factor(STNAME)Utah
## factor(STNAME)Vermont            ***
## factor(STNAME)Virginia           **
## factor(STNAME)Washington
## factor(STNAME)West Virginia
## factor(STNAME)Wisconsin          ***
## factor(STNAME)Wyoming
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 25.72 on 3089 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.4521, Adjusted R-squared:  0.4431
## F-statistic: 49.98 on 51 and 3089 DF,  p-value: < 2.2e-16
```

```
ggplot(rawdata, aes(MeanMedianIncome,CasePer100k))+geom_point()+geom_abline(intercept=coef(regr3)[1],slope=coef(regr3)[2],lwd=1, col="red")
```



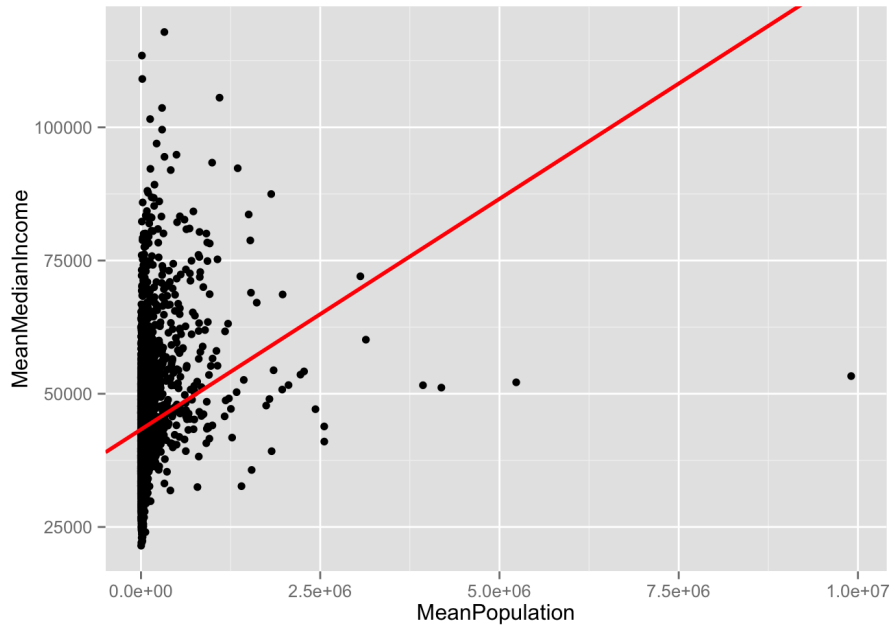
APPENDIX IV

```
regr4<-lm(MeanMedianIncome ~ MeanPopulation, data=rawdata)
summary(regr4)
```

```
##
## Call:
## lm(formula = MeanMedianIncome ~ MeanPopulation, data = rawdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -75728  -7092  -1747   4921  71760
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.330e+04  2.012e+02   215.2   <2e-16 ***
## MeanPopulation 8.656e-03  6.054e-04    14.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10760 on 3139 degrees of freedom
##   (2 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.06114,    Adjusted R-squared:  0.06084
## F-statistic: 204.4 on 1 and 3139 DF,  p-value: < 2.2e-16
```

```
ggplot(rawdata, aes(MeanPopulation,MeanMedianIncome))+geom_point()
+geom_abline(intercept=coef(regr4)[1],slope=coef(regr4)[2],lwd=1, col="red")
```



Appendix VI

Summaries of Linear Regressions:

Linear Regression- Education using data for all 50 United States:

```
================================================================================================================
=====================================================
  Regression of lessThanHighSchoolDiploma  Regression of highSchooDiplomaOnly  Regression of
someCollegeOrAssociateDegree  Regression of bachelorDegreeOrHigher
----------------------------------------------------------------------------------------------------------------
----------------------------------------
(Intercept)                0.03 ***                 0.05 ***                 0.04 ***                 0.02 ***
                          (0.00)                   (0.00)                   (0.00)                   (0.00)
aggregateCasesPerOneHundredThousand  -0.35 ***                -1.01 ***                -0.24 **
1.36 ***
                          (0.08)                   (0.11)                   (0.07)                   (0.12)
----------------------------------------------------------------------------------------------------------------
----------------------------------------
Adj. R^2                   0.40                     0.28                     0.38                     0.25
Num. obs.                  3142                     3142                     3142                     3142
```

==========
==========
*** p < 0.001, ** p < 0.01, * p < 0.05

Coefficients for Education using data for all 50 United States:

Less than a High School Diploma:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0296943	0.0008782	33.813	< 2e-16 ***

High School Diploma Only:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0504220	0.0011830	42.621	< 2e-16 ***

Some College or Associate's:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0398529	0.0008130	49.018	< 2e-16 ***

Bachelor Degree or Higher:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0230700	0.0012667	18.213	< 2e-16 ***

Multiple Linear Regression- Education and Median Income using data for all 50 States:

	Regression of lessThanHighSchoolDiploma	Regression of highSchooDiplomaOnly with medianIncome	Regression of someCollegeOrAssociateDegree with medianIncome	Regression of bachelorDegreeOrHigher with medianIncome
(Intercept)	0.23 *	0.26 *	0.25 *	0.25 *
	(0.10)	(0.10)	(0.10)	(0.10)
aggregateCasesPerOneHundredThousand	-13.25 **	-13.74 **	-13.21 **	-12.33 **
	(4.10)	(4.13)	(4.10)	(4.09)
Adj. R^2	0.09	0.10	0.09	0.08
Num. obs.	91	91	91	91

==========
==========
*** p < 0.001, ** p < 0.01, * p < 0.05

Coefficients for Education and Median Income using data for all 50 United States:

Less than a High School Diploma:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.27781	0.31477	0.883	0.377536	

High School Diploma Only:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	142.4736	24.5775	5.797	7.44e-09	***

Some College or Associate's:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	5.706e+01	2.625e+01	2.174	0.02979	*

Bachelor Degree or Higher:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-71.3873	20.2677	-3.522	0.000434	***

Linear Regression- Education for Endemic States MA, PA, CT:

	Regression of lessThanHighSchoolDiploma In Endemic Regions	Regression of highSchooDiplomaOnly in Endemic Regions	Regression of someCollegeOrAssociateDegree In Endemic Regions	Regression of bachelorDegreeOrHigher In Endemic Regions
(Intercept)	0.02 ***	0.05 ***	0.04 ***	0.04 ***
	(0.00)	(0.00)	(0.00)	(0.00)
aggregateCasesPerOneHundredThousand	-0.17 *	-0.66 ***	-0.13	0.75 ***
	(0.07)	(0.14)	(0.07)	(0.15)
Adj. R^2	0.15	0.60	0.12	0.56
Num. obs.	91	91	91	91

*** p < 0.001, ** p < 0.01, * p < 0.05

Coefficients-Education for Endemic States MA, PA, CT:

Less than a High School Diploma:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.0156117	0.0016771	9.309	1.05e-14	***

High School Diploma Only:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.048147	0.003627	13.273	< 2e-16	***

Some College or Associate's:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.0377925	0.0017041	22.178	<2e-16	***

Bachelor Degree or Higher:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.037941	0.003724	10.188	< 2e-16 ***

Multiple Linear Regression- Education and Median Income for Endemic States MA, PA, CT:

	Regression of lessThanHighSchoolDiploma In Endemic Regions	Regression of highSchooDiplomaOnly in Endemic Regions	Regression of someCollegeOrAssociateDegree In Endemic Regions	Regression of bachelorDegreeOrHigher In Endemic Regions
(Intercept)	0.23 *	0.26 *	0.25 *	0.25 *
	(0.10)	(0.10)	(0.10)	(0.10)
aggregateCasesPerOneHundredThousand	-13.25 **	-13.74 **	-13.21 **	-12.33 **
	(4.10)	(4.13)	(4.10)	(4.09)
Adj. R^2	0.09	0.10	0.09	0.08
Num. obs.	91	91	91	91

*** p < 0.001, ** p < 0.01, * p < 0.05

Multiple Linear Regression- Education and Median Income for Endemic States MA, PA, CT:

Less than a High School Diploma:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.23234	0.10362	2.242	0.02749 *

High School Diploma Only:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.26488	0.10452	2.534	0.0131 *

Some College or Associate's:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.254520	0.103595	2.457	0.01600 *

Bachelor Degree or Higher

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.547e-01	1.034e-01	2.463	0.01575 *

Appendix VII

Code for education analysis:

```
library(texreg)
data<-read.csv("finalData.csv")
```

```
#Check individual education columns to test how good of a fit they are
edFit1 <- lm( lessHSPercent ~ aggregateCasesPerOneHundredThousand + factor(STNAME), data)

edFit2 <- lm( hsOnlyPercent ~ aggregateCasesPerOneHundredThousand  + factor(STNAME), data)

edFit3<-  lm( someCollegePercent ~ aggregateCasesPerOneHundredThousand  + factor(STNAME),
data)

edFit4<-  lm( bachelorPercent ~ aggregateCasesPerOneHundredThousand  + factor(STNAME), data)


#Plot education regressions
#edFit1 plot
plot(data[,22], data[,32], ann=FALSE)
abline(edFit1)
abline(edFit1, lwd=3)
abline(edFit1, lwd=3, col="red")
title(main = "Less Than a High School Education",sub= NULL, xlab = "Cases of Lyme Disease By
County", ylab = "% of Persons with Less Than a High School Education by County")

screenreg( list(edFit1), omit.coef="factor", custom.model.names=c("Regression of Ed1"),
include.rsquared=FALSE, include.rmse=FALSE)

#edFit2 plot
plot(data[,22], data[,34], ann=FALSE)
abline(edFit2)
abline(edFit2, lwd=3)
abline(edFit2, lwd=3, col="red")
title(main = "High School Diploma Only",sub= NULL, xlab = "Cases of Lyme Disease By County", ylab =
"% of Persons with High School Diploma Only")

screenreg( list(edFit2), omit.coef="factor", custom.model.names=c("Regression of Ed1"),
include.rsquared=FALSE, include.rmse=FALSE)


#edFit3  plot
plot(data[,22], data[,36], ann=FALSE)
abline(edFit3)
abline(edFit3, lwd=3)
abline(edFit3, lwd=3, col="red")
title(main = "Some College or Associate's Degree",sub= NULL, xlab = "Cases of Lyme Disease By
County", ylab = "% of Persons with Some College or Associate's Degree")

screenreg( list(edFit3), omit.coef="factor", custom.model.names=c("Regression of Ed1"),
include.rsquared=FALSE, include.rmse=FALSE)


#edFit4 plot
plot(data[,22], data[,38], ann=FALSE)
abline(edFit4)
abline(edFit4, lwd=3)
abline(edFit4, lwd=3, col="red")
```

```
title(main = "Bachelor's Degree or Higher",sub= NULL, xlab = "Cases of Lyme Disease By County",
ylab = "% of Persons with Bachelor's Degree or Higher")

screenreg( list(edFit4), omit.coef="factor", custom.model.names=c("Regression of Ed1"),
include.rsquared=FALSE, include.rmse=FALSE)
```

Code for multiple linear regression Education + Median Income:

```
library(texreg)
data<-read.csv("finalData.csv")
fit1 <- lm( aggregateCases ~ lessHighSchoolDiploma + aggregateMedianHouseInc + factor(STNAME),
data)

fit2 <- lm( aggregateCases ~ highSchoolDiplomaOnly + aggregateMedianHouseInc + factor(STNAME),
data)

fit3<-  lm( aggregateCases ~ someCollegeOrAssociateDegree + aggregateMedianHouseInc +
factor(STNAME), data)

fit4<-  lm( aggregateCases ~ bachelorDegreeOrHigher + aggregateMedianHouseInc +
 factor(STNAME), data)

combinedEducation<-screenreg( list(edFit1, edFit2, edFit3, edFit4), omit.coef="factor",
custom.model.names=c("Regression of lessThanHighSchoolDiploma with medianIncome",
"Regression of highSchooDiplomaOnly with medianIncome", "Regression of
someCollegeOrAssociateDegree with medianIncome", "Regression of bachelorDegreeOrHigher with
medianIncome"), include.rsquared=FALSE, include.rmse=FALSE)
```

Code for linear regression and plots in endemic regions:

```
library(texreg)
data<-read.csv("finalDataEdemicAreas.csv")

#Check individual education columns to test how good of a fit they are
edFit1 <- lm( lessHSPercent ~ aggregateCasesPerOneHundredThousand + factor(STNAME), data)

edFit2 <- lm( hsOnlyPercent ~ aggregateCasesPerOneHundredThousand  + factor(STNAME), data)

edFit3<-  lm( someCollegePercent ~ aggregateCasesPerOneHundredThousand  + factor(STNAME),
data)

edFit4<-  lm( bachelorPercent ~ aggregateCasesPerOneHundredThousand  + factor(STNAME), data)




#Plot education regressions for endemic regions

#edFit1
plot(data[,22], data[,32], ann=FALSE)
abline(edFit1)
abline(edFit1, lwd=3)
abline(edFit1, lwd=3, col="red")
```

```
title(main = "Less Than a High School Education",sub= NULL, xlab = "Cases of Lyme Disease By County", ylab = "% of Persons with Less Than a High School Education by County")

screenreg( list(edFit1), omit.coef="factor", custom.model.names=c("Regression of Ed1"), include.rsquared=FALSE, include.rmse=FALSE)

#edFit2
plot(data[,22], data[,34], ann=FALSE)
abline(edFit2)
abline(edFit2, lwd=3)
abline(edFit2, lwd=3, col="red")
title(main = "High School Diploma Only",sub= NULL, xlab = "Cases of Lyme Disease By County", ylab = "% of Persons with High School Diploma Only")


screenreg( list(edFit2), omit.coef="factor", custom.model.names=c("Regression of Ed1"), include.rsquared=FALSE, include.rmse=FALSE)


#edFit3
plot(data[,22], data[,36], ann=FALSE)
abline(edFit3)
abline(edFit3, lwd=3)
abline(edFit3, lwd=3, col="red")
title(main = "Some College or Associate's Degree",sub= NULL, xlab = "Cases of Lyme Disease By County", ylab = "% of Persons with Some College or Associate's Degree")


screenreg( list(edFit3), omit.coef="factor", custom.model.names=c("Regression of Ed1"), include.rsquared=FALSE, include.rmse=FALSE)


#edFit4
plot(data[,22], data[,38], ann=FALSE)
abline(edFit4)
abline(edFit4, lwd=3)
abline(edFit4, lwd=3, col="red")
title(main = "Bachelor's Degree or Higher",sub= NULL, xlab = "Cases of Lyme Disease By County", ylab = "% of Persons with Bachelor's Degree or Higher")


screenreg( list(edFit4), omit.coef="factor", custom.model.names=c("Regression of Ed1"), include.rsquared=FALSE, include.rmse=FALSE)

combinedEducation<-screenreg( list(edFit1, edFit2, edFit3, edFit4), omit.coef="factor", custom.model.names=c("Regression of lessThanHighSchoolDiploma In Endemic Regions", "Regression of highSchooDiplomaOnly in Endemic Regions", "Regression of someCollegeOrAssociateDegree In Endemic Regions", "Regression of bachelorDegreeOrHigher In Endemic Regions"), include.rsquared=FALSE, include.rmse=FALSE)
```