Diabetes_Poverty

Bret Harvestine

12/19/2016

Is poverty becoming a rising cause in the prevalence of Type II Diabetes?

Introduction

This paper constructs statistical models to fit and perform linear regression to test the relationship between poverty and the prevalence of type II diabetes. The prevalence of type II diabetes (2013) will be the response variable and will be tested against the treatment variable, poverty rates by county (2013). Scholarly studies are emerging that show living in low-income and poverty stricken areas is a leading cause of type II diabetes. Additionally, once they do develop the disease, they are much more likely to suffer complications such as amputations, blindness, or cardiovascular disease. There are numerous factors in poverty that influence this situation. People who have low income have a higher chance of not being able to afford health insurance, and they have low resources to healthy foods and time for fitness. Consequently, the lack of health resources could negatively impact healthcare costs: people of poor health history are more likely to have more expensive healthcare costs than a person of good health. A combination of all these factors, in addition to having poor access to healthy food and grocery stores are resulting causes of a higher prevalence of type II diabetes diagnoses. Therefore, the linear models performed in this paper will include additional treatments to be tested against the response variable: median household income by county, uninsured rates by county, access to grocery stores by county, adult leisure time spent being physically inactive by county, percentage of obese adults by county, percent of households recieving food stamp assistance by county, percentage of adult smokers by county, and percentage of households with low food security by county. These additional treatments are several different representations of living in poverty, which work together to exemplify the rising issue that people in the United States that are living in poverty, have a greater likelihood of being diagnosed with type II diabetes.

Methods

Data was carefully collected from the U.S. Census Bureau, healthdata.gov, United States Department of Agriculture Economic Research Service, and the Centers for Disease Control and Prevention.

Data for these linear regression models are specified to the year 2013. Data was observed under histogram plots in order to identify outliers. If outliers were discovered, subsets of data were created to clean the data for possible bias in the regression models.

```
## Loading required package: blscrapeR
```

```
## Loading required package: choroplethr
```

```
## Loading required package: acs
```

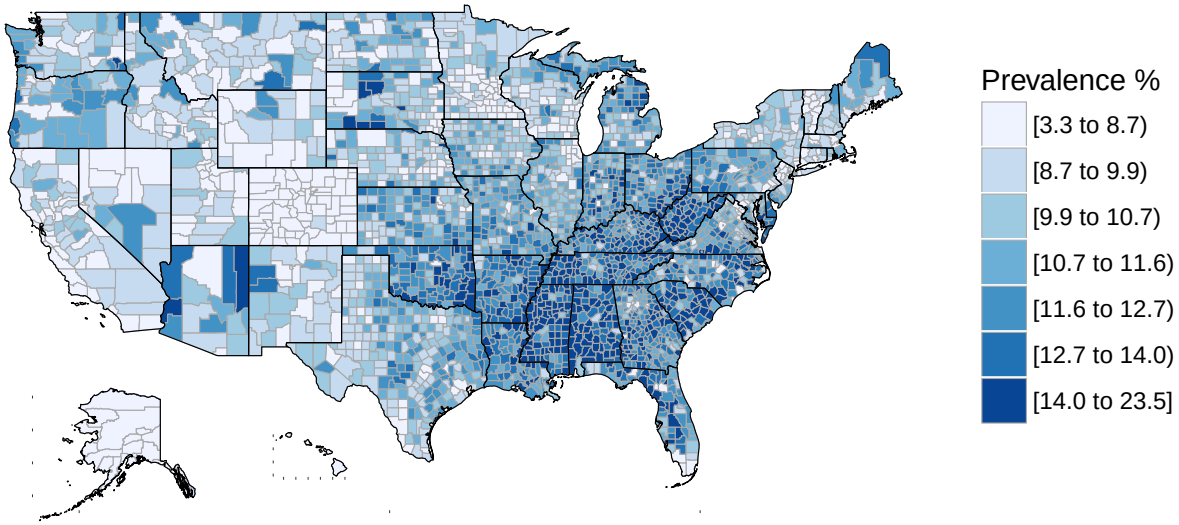
```
## Loading required package: stringr
```

```
## Loading required package: plyr

## Loading required package: XML

##
## Attaching package: 'acs'

## The following object is masked from 'package:base':
##
##     apply

## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:acs':
##
##     combine

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

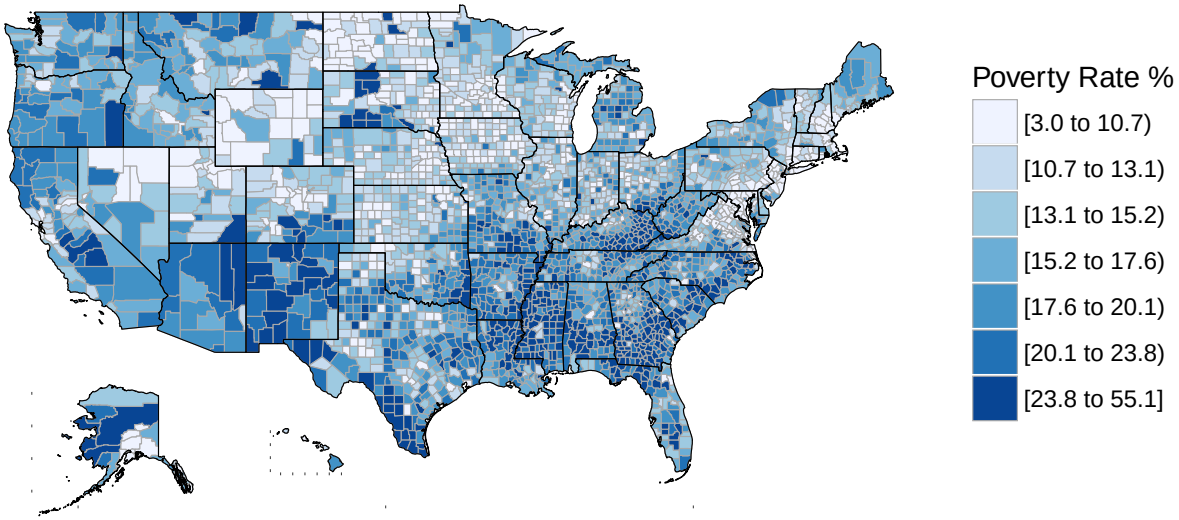
Load in Data

cbind() function was used to combine data and prepare to merge into one dataframe. Data was bound under two columns, region and value, and used with the choroplethr pagackage to map data by county in order to observe and compare national or regional trends.

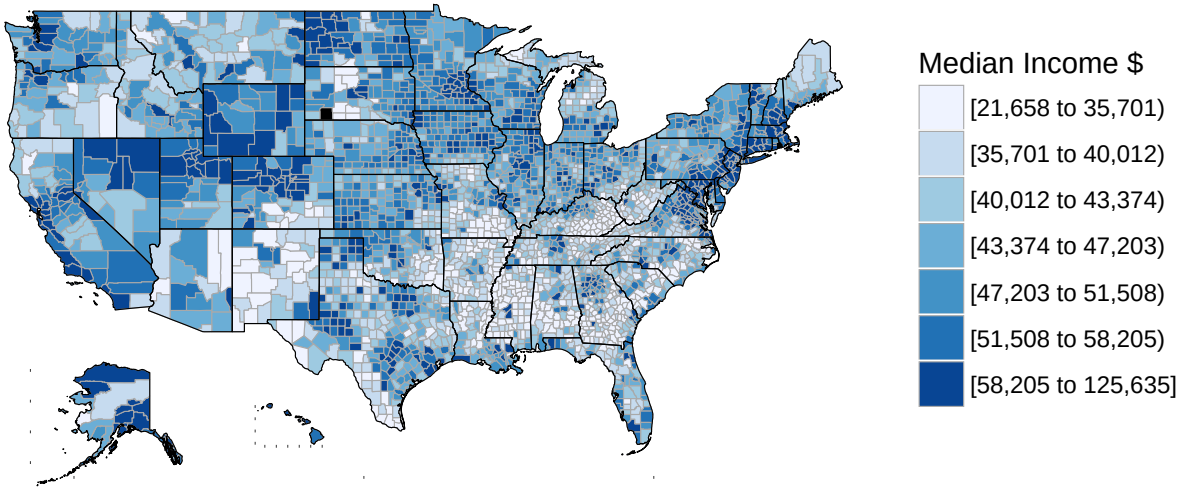
Prevalence of Diabetes by County 2013



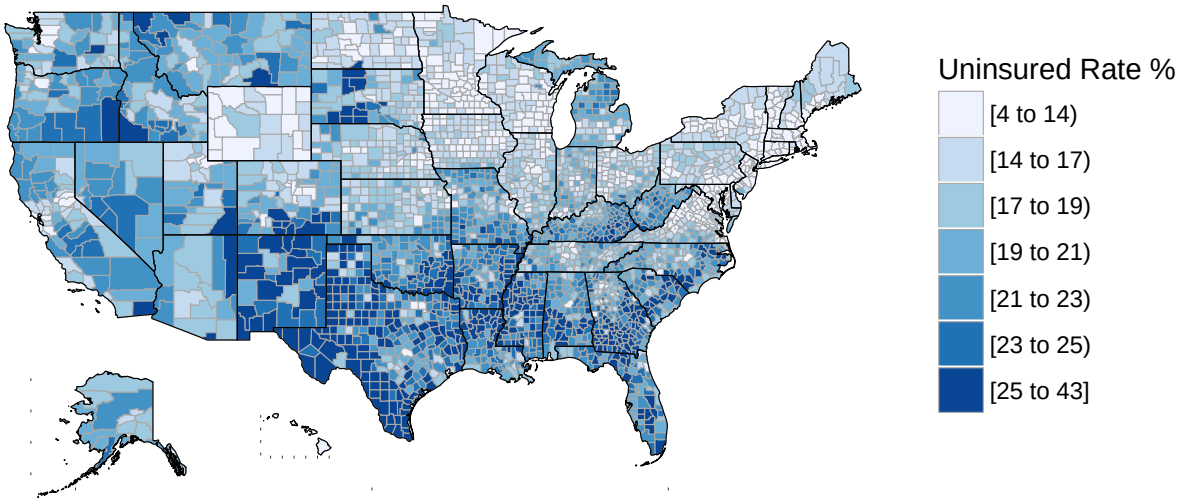
Poverty Rate by County 2013



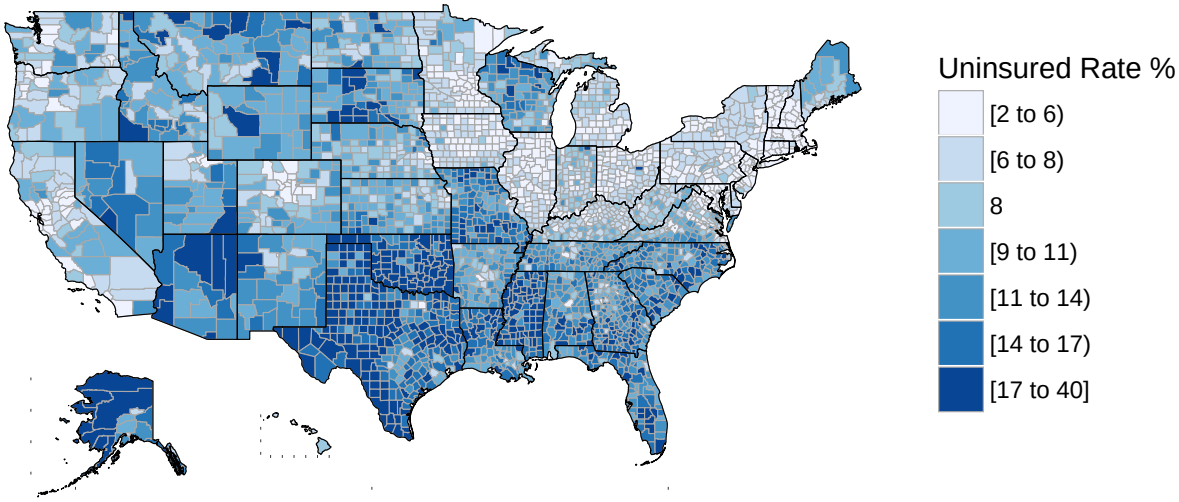
Median Household Income by County 2013



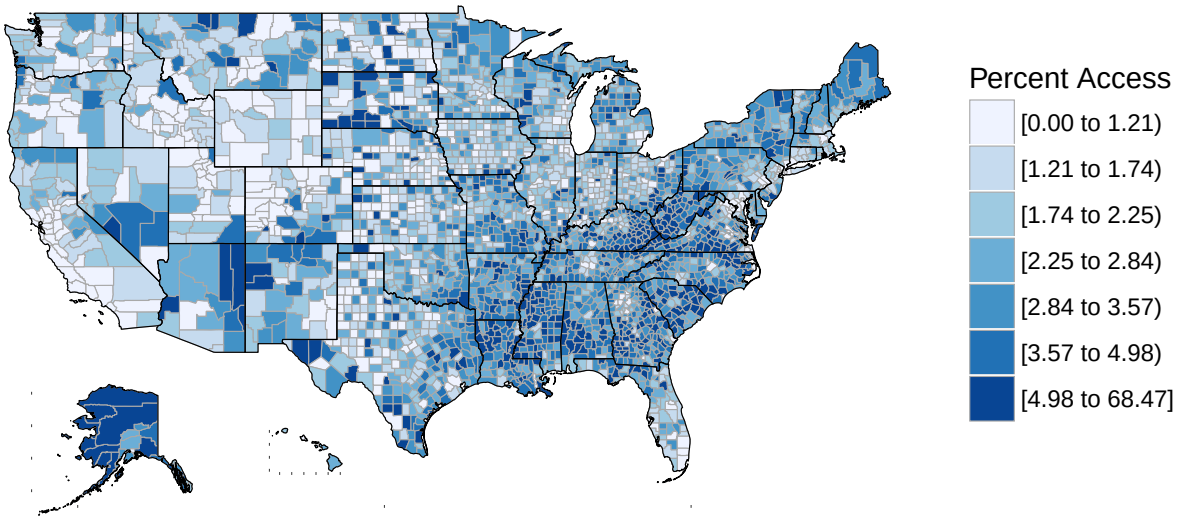
Uninsured Rate by County 2013



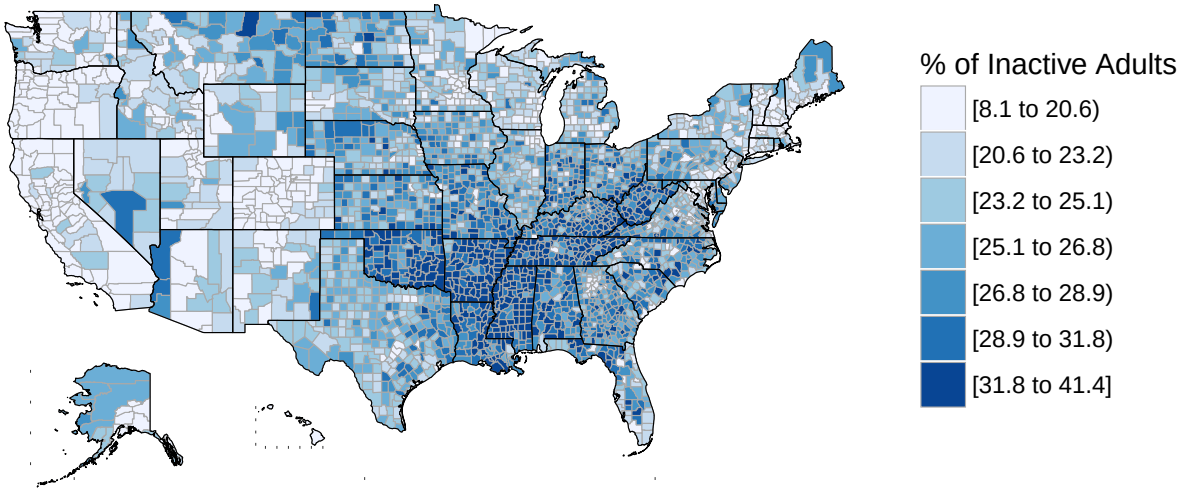
Uninsured Rate by County 2016



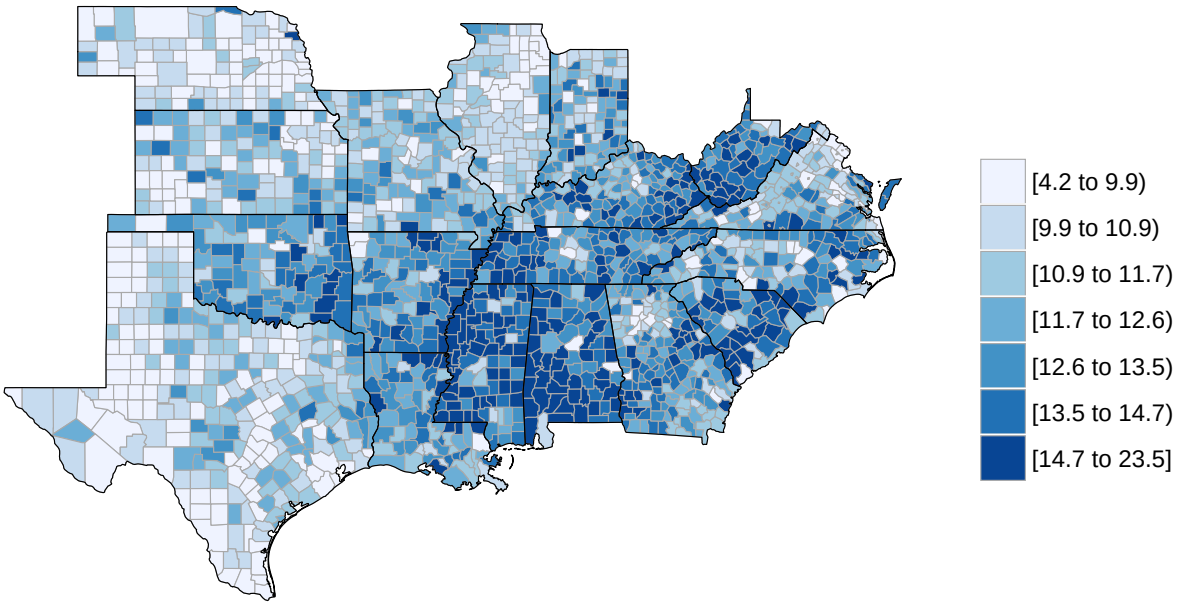
Access to Grocery Stores by County 2013



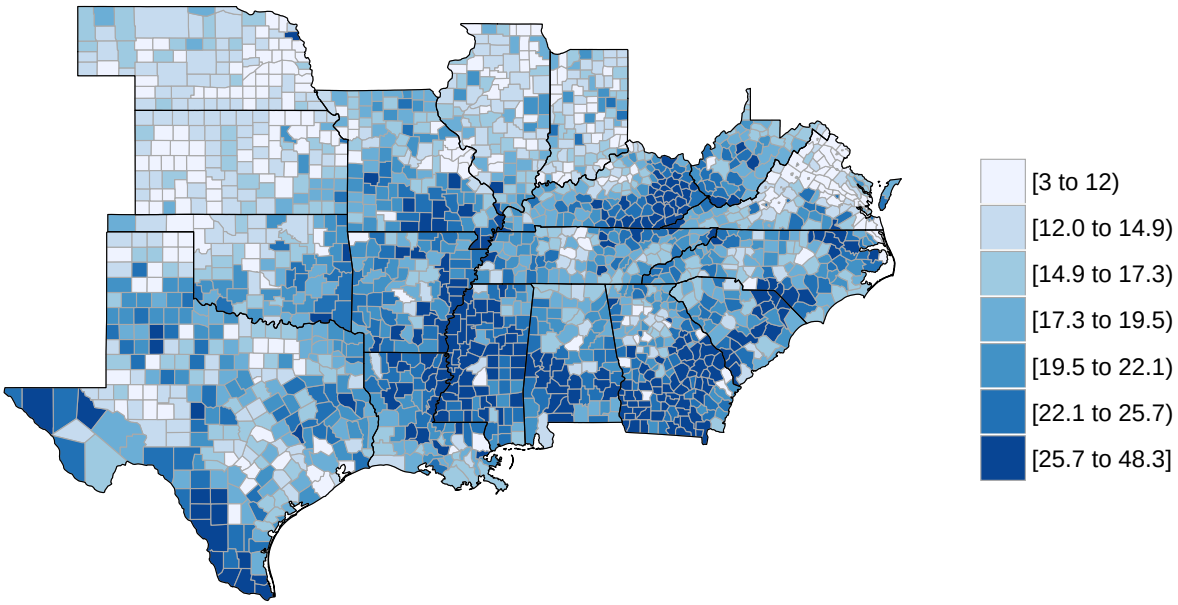
Leisure Time Spent Physically Inactive by County 2013



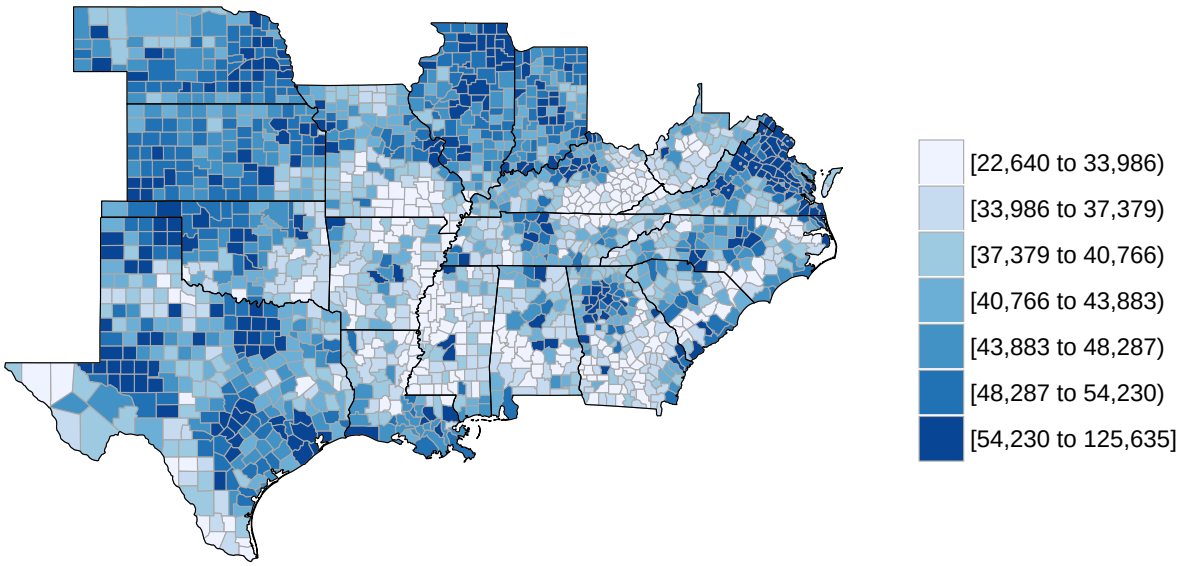
Diabetes Prevalence (%)



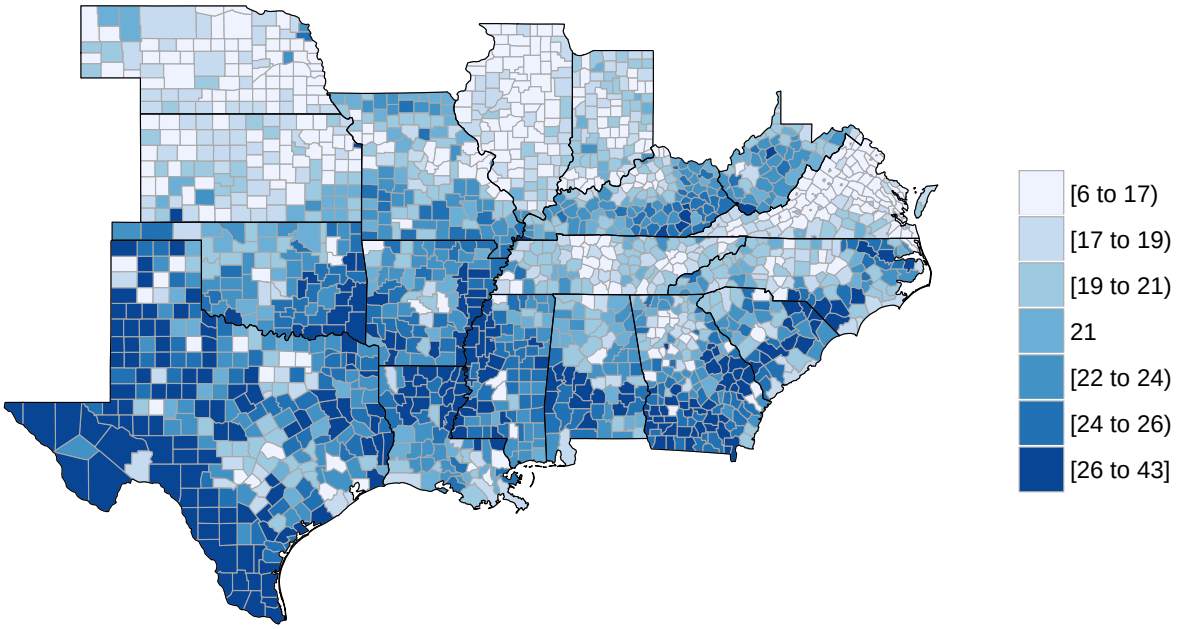
Poverty Rate (%)

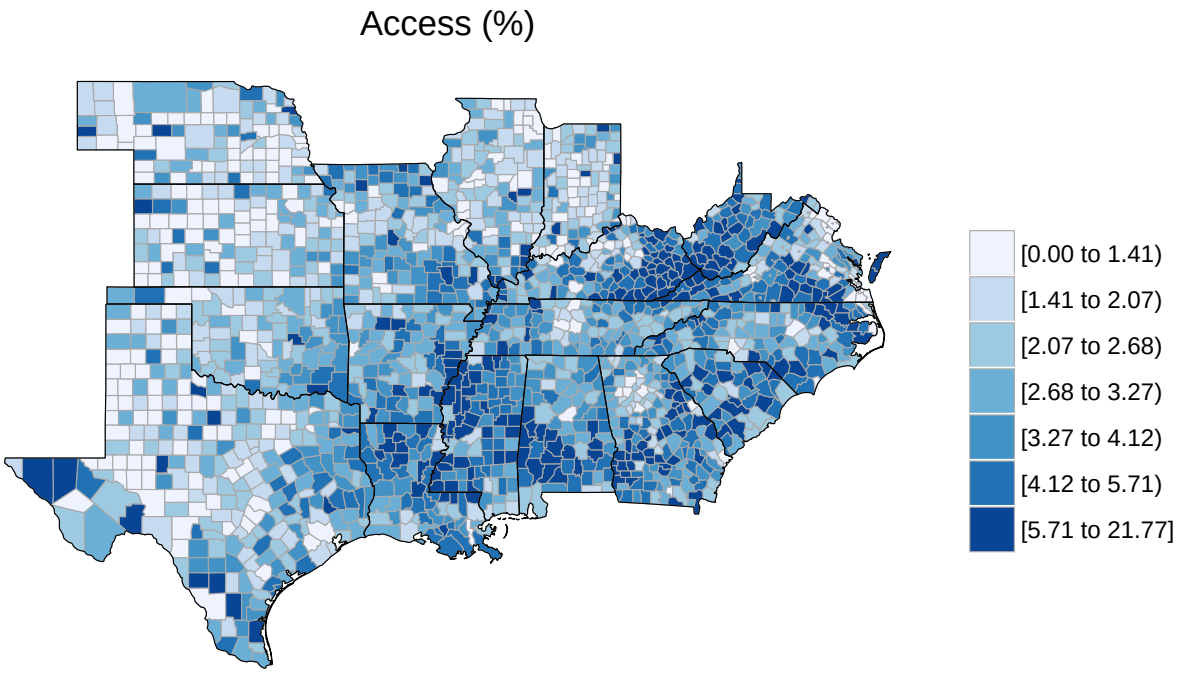


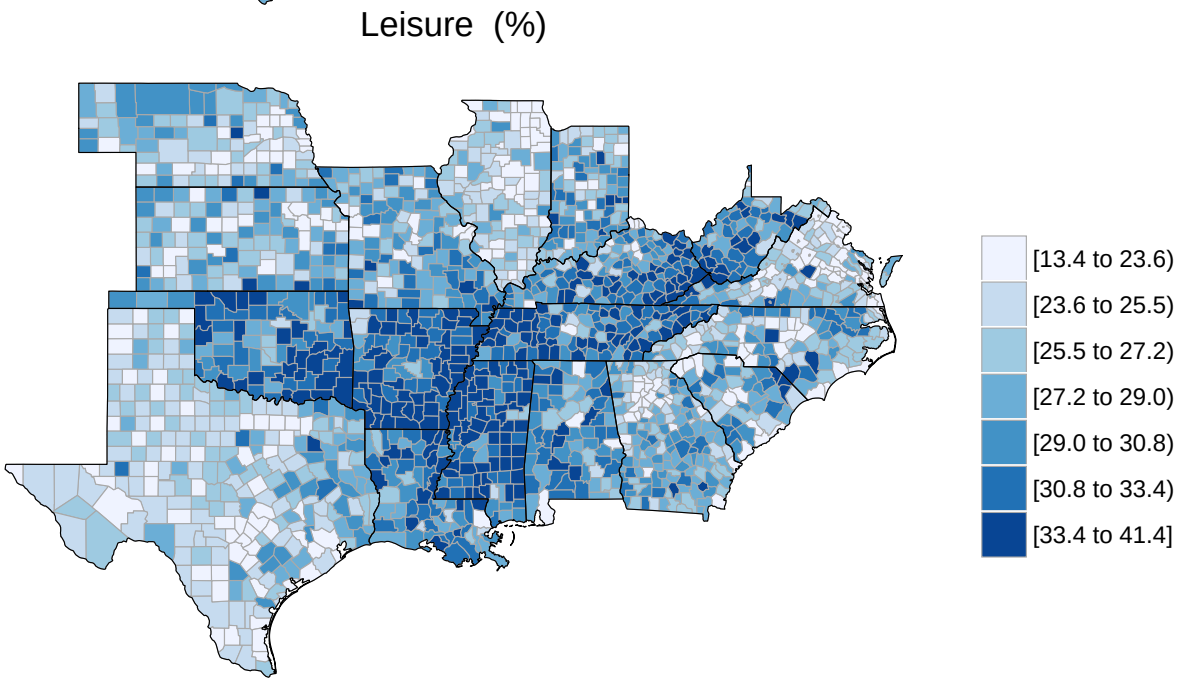
Median Household Income



Uninsured Rate (%)







Notice the switch? Initial observations using choroplethr map analysis allude to a noticable correlation between the prevalence of type II diabetes, median household income, and leisure time. There are also similar patterns displayed in the independent variable, access to healthy food and grocery stores by county. Counties that report a high prevalence in type II diabetes also reported low median household income, high amount of leisure time spent inactive, and low access to grocery stores.

Data was merged into one dataframe and prepared for linear regression. The final data frame will be referenced as final.final in the linear models.

```
##   FIPS uninsured13 uninsured16 prevalence income poverty   access leisure
## 1 1001          17           8       13.0  54366    13.5 3.284786    28.6
## 2 1003          17           8       10.4  49626    14.2 2.147827    22.3
## 3 1005          24          15       18.4  34971    28.2 4.135869    31.8
## 4 1007          20          10       14.8  39546    23.1 3.458580    33.9
## 5 1009          19          10       14.1  45567    17.2 3.269380    28.0
## 6 1011          28          17       19.6  26580    35.2 7.285561    31.7
##   assistance fastfood insecurity obese smokers
## 1   14.42302 52.17391        6.8  32.4    23.5
## 2   14.42302 35.44304        6.8  32.4    22.6
## 3   14.42302 55.88235        6.8  32.4    23.5
## 4   14.42302 45.45455        6.8  32.4    32.8
## 5   14.42302 56.41026        6.8  32.4    22.0
## 6   14.42302 50.00000        6.8  32.4      NA
```

Linear Regressions:

Response: Diabetes Prevalence

Treatment: Poverty Rate; Access and Proximity to Grocery Stores; Uninsured Rate; Median Household Income; Leisure Time spent Physically Inactive; Food Stamp Assistance; Percent Obese Adults; Percent Adult Smokers; Percent Fast Food Restaurants

```
#simple linear models

lm.p = lm(prevalence ~ poverty, data = final)
summary(lm.p)
```

```
##
## Call:
## lm(formula = prevalence ~ poverty, data = final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3163  -1.2772   0.0481   1.3602   8.3954
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.873138   0.106356   74.03   <2e-16 ***
## poverty     0.195444   0.005777   33.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.128 on 3148 degrees of freedom
## Multiple R-squared:  0.2667, Adjusted R-squared:  0.2664
## F-statistic:  1145 on 1 and 3148 DF,  p-value: < 2.2e-16
```

```
lm.i = lm(prevalence ~ income, data = final)
summary(lm.i)
```

```
##
## Call:
## lm(formula = prevalence ~ income, data = final)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5359 -1.2927 -0.0093  1.3177 10.3405
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.675e+01  1.462e-01  114.53   <2e-16 ***
## income      -1.169e-04  3.005e-06  -38.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.042 on 3148 degrees of freedom
## Multiple R-squared:  0.3249, Adjusted R-squared:  0.3247
## F-statistic:  1515 on 1 and 3148 DF,  p-value: < 2.2e-16
```

```
lm.un13 = lm(prevalence ~ uninsured13, data = final)
summary(lm.un13)
```

```
##
## Call:
## lm(formula = prevalence ~ uninsured13, data = final)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.998 -1.314  0.060  1.432 10.131
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.301853   0.150837   48.41   <2e-16 ***
## uninsured13 0.209212   0.007737   27.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.239 on 3148 degrees of freedom
## Multiple R-squared:  0.1885, Adjusted R-squared:  0.1882
## F-statistic: 731.1 on 1 and 3148 DF,  p-value: < 2.2e-16
```

```
lm.a = lm(prevalence ~ access, data = final)
summary(lm.a)
```

```
##
## Call:
## lm(formula = prevalence ~ access, data = final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.2342  -1.5376  -0.0584   1.5578  10.2750
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.57822    0.05992  176.55   <2e-16 ***
## access       0.20821    0.01333   15.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.394 on 3148 degrees of freedom
## Multiple R-squared:  0.07188,    Adjusted R-squared:  0.07158
## F-statistic: 243.8 on 1 and 3148 DF,  p-value: < 2.2e-16
```

```
lm.l = lm(prevalence ~ leisure, data = final)
summary(lm.l)
```

```
##
## Call:
## lm(formula = prevalence ~ leisure, data = final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1450 -1.1067 -0.0678  0.9465 10.8701
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.04701    0.15118   13.54   <2e-16 ***
## leisure      0.35394    0.00571   61.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.668 on 3148 degrees of freedom
## Multiple R-squared:  0.5496, Adjusted R-squared:  0.5495
## F-statistic:  3842 on 1 and 3148 DF,  p-value: < 2.2e-16
```

```
#multiple linear regression

mult.lm = lm(prevalence ~  access + assistance + fastfood+  poverty + leisure + obese+ insecurity + smol
summary(mult.lm)
```

```
##
## Call:
## lm(formula = prevalence ~ access + assistance + fastfood + poverty +
##     leisure + obese + insecurity + smokers, data = final.final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2175 -0.9742 -0.0386  0.9190  9.0241
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.230926   0.379926  -0.608 0.543364
## access       0.041553   0.013420   3.096 0.001982 **
## assistance   0.116771   0.015929   7.331  3.1e-13 ***
## fastfood     0.002727   0.002506   1.088 0.276717
## poverty      0.059324   0.006418   9.243  < 2e-16 ***
## leisure      0.247976   0.008682  28.562  < 2e-16 ***
## obese        0.057453   0.016273   3.531 0.000422 ***
## insecurity   0.067798   0.037526   1.807 0.070933 .
## smokers      0.017459   0.006342   2.753 0.005953 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.525 on 2442 degrees of freedom
```

```
##   (696 observations deleted due to missingness)
## Multiple R-squared:  0.6507, Adjusted R-squared:  0.6496
## F-statistic: 568.7 on 8 and 2442 DF,  p-value: < 2.2e-16
```

```
#randomForest package regression analysis
library(randomForest)
```

```
## randomForest 4.6-12
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

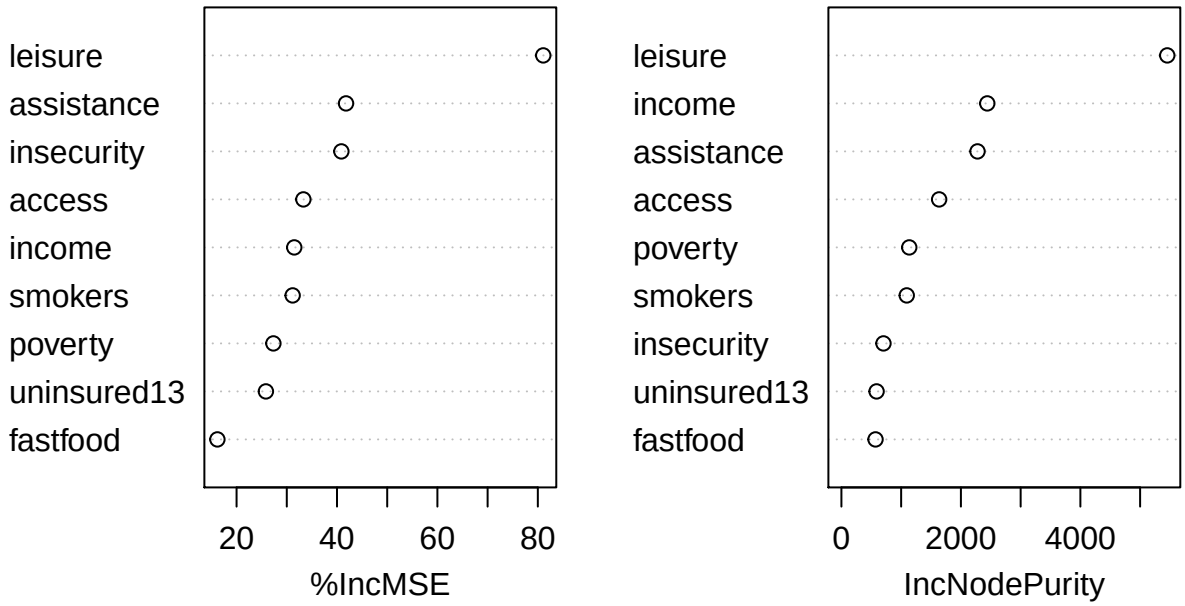
```
## The following object is masked from 'package:acs':
##
##     combine
```

```
rf.lm = randomForest(prevalence ~ access + assistance + fastfood + income + insecurity + leisure + pover
print(rf.lm)
```

```
##
## Call:
##  randomForest(formula = prevalence ~ access + assistance + fastfood +      income + insecurity + leis
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 1.724161
##                     % Var explained: 74.02
##   Bias correction applied:
##   Intercept:  -0.01166764
##       Slope:  1.045461
```

```
#Plot randomForest regression model to find importance
varImpPlot(rf.lm)
```

rf.lm




```
#add interactions

smoke.lm = lm(prevalence ~ access + fastfood +smokers+ assistance*poverty + uninsured13*income + poverty
summary(smoke.lm)
```

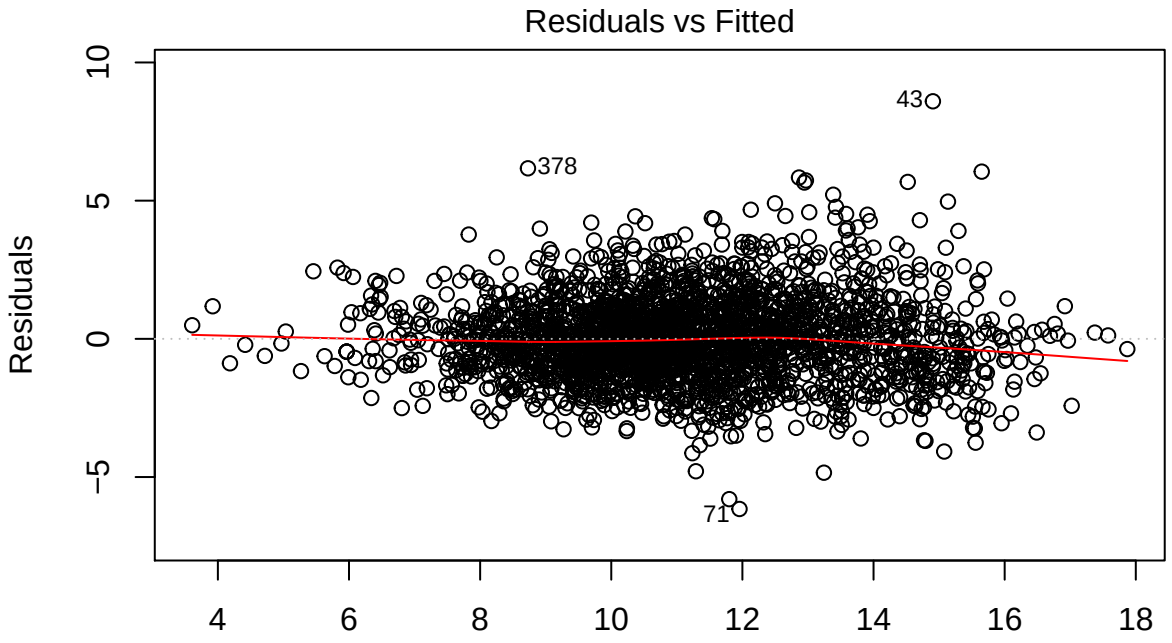
```
##
## Call:
## lm(formula = prevalence ~ access + fastfood + smokers + assistance *
##     poverty + uninsured13 * income + poverty * leisure, data = final.final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2065 -0.9889 -0.0657  0.9180  9.3524
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.846e+00  8.192e-01   4.696 2.80e-06 ***
## access               3.226e-02  1.321e-02   2.443 0.014648 *
## fastfood             6.267e-03  2.545e-03   2.463 0.013854 *
## smokers              1.121e-02  6.330e-03   1.770 0.076835 .
## assistance           2.579e-01  3.365e-02   7.665 2.57e-14 ***
## poverty             -4.291e-02  3.065e-02  -1.400 0.161646
## uninsured13          1.237e-01  2.791e-02   4.431 9.77e-06 ***
## income               1.851e-06  7.773e-06   0.238 0.811798
## leisure              1.679e-01  2.194e-02   7.653 2.82e-14 ***
## assistance:poverty -6.083e-03  1.934e-03  -3.146 0.001677 **
## uninsured13:income -3.531e-06  5.401e-07  -6.538 7.55e-11 ***
## poverty:leisure      4.358e-03  1.187e-03   3.672 0.000246 ***
## ---
```

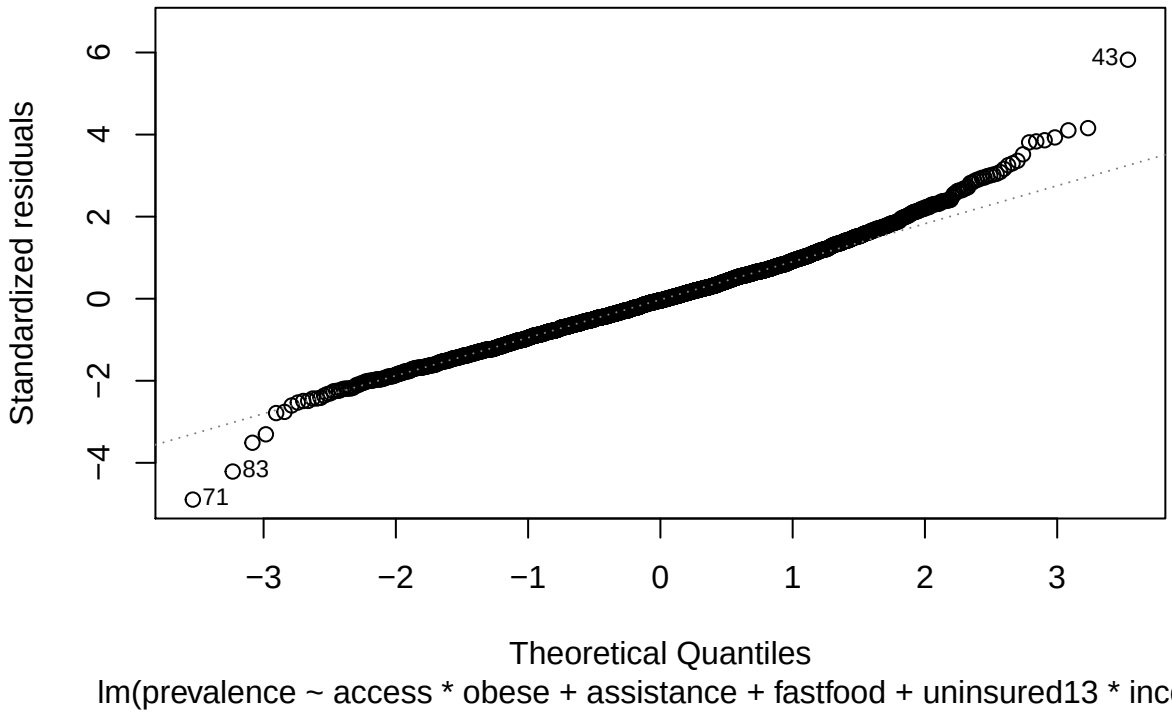
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.496 on 2439 degrees of freedom
##   (696 observations deleted due to missingness)
## Multiple R-squared:  0.6646, Adjusted R-squared:  0.6631
## F-statistic: 439.3 on 11 and 2439 DF,  p-value: < 2.2e-16
```

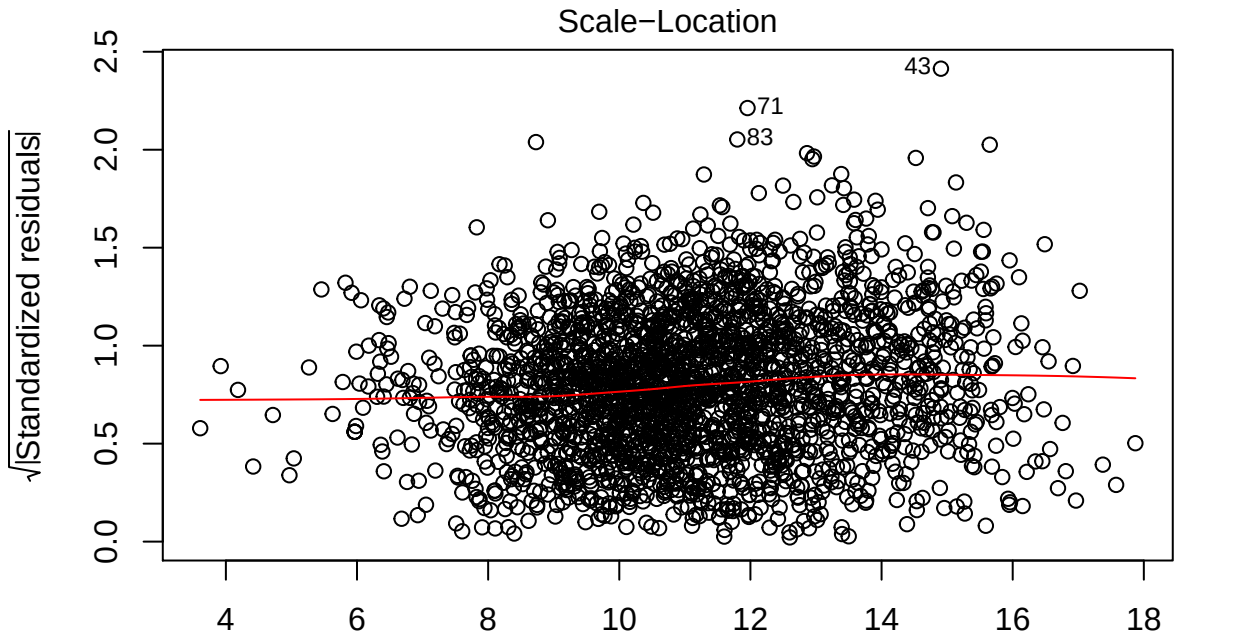
```
#final linear model with smoking
final.lm = lm(prevalence ~ access*obese + assistance + fastfood + uninsured13*income + leisure*smokers +
summary(final.lm)
```

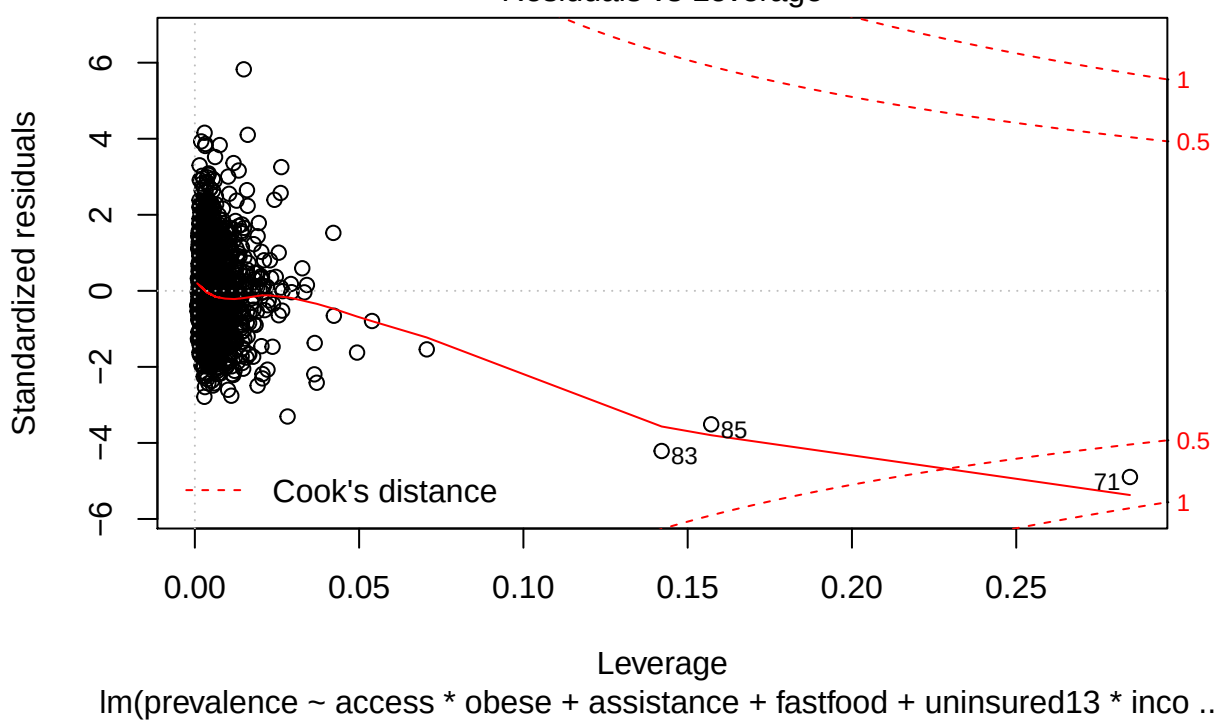
```
##
## Call:
## lm(formula = prevalence ~ access * obese + assistance + fastfood +
##     uninsured13 * income + leisure * smokers + poverty, data = final.final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1558 -0.9614 -0.0611  0.8910  8.5948
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.722e-01  8.628e-01   0.663  0.50729
## access              -5.520e-01  1.433e-01  -3.852  0.00012 ***
## obese                4.130e-03  2.035e-02   0.203  0.83918
## assistance           1.256e-01  1.455e-02   8.634  < 2e-16 ***
## fastfood             7.234e-03  2.535e-03   2.854  0.00435 **
## uninsured13          1.350e-01  2.672e-02   5.054 4.65e-07 ***
## income               1.138e-05  6.887e-06   1.652  0.09871 .
## leisure              3.204e-01  2.192e-02  14.612  < 2e-16 ***
## smokers              1.441e-01  2.730e-02   5.277 1.43e-07 ***
## poverty              7.908e-03  1.146e-02   0.690  0.49035
## access:obese         1.994e-02  4.851e-03   4.109 4.10e-05 ***
## uninsured13:income -3.823e-06  5.033e-07  -7.597 4.29e-14 ***
## leisure:smokers    -4.855e-03  1.000e-03  -4.855 1.28e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.487 on 2438 degrees of freedom
##   (696 observations deleted due to missingness)
## Multiple R-squared:  0.6687, Adjusted R-squared:  0.6671
## F-statistic: 410.1 on 12 and 2438 DF,  p-value: < 2.2e-16
```

```
#Plot linear models

plot(final.lm)
```

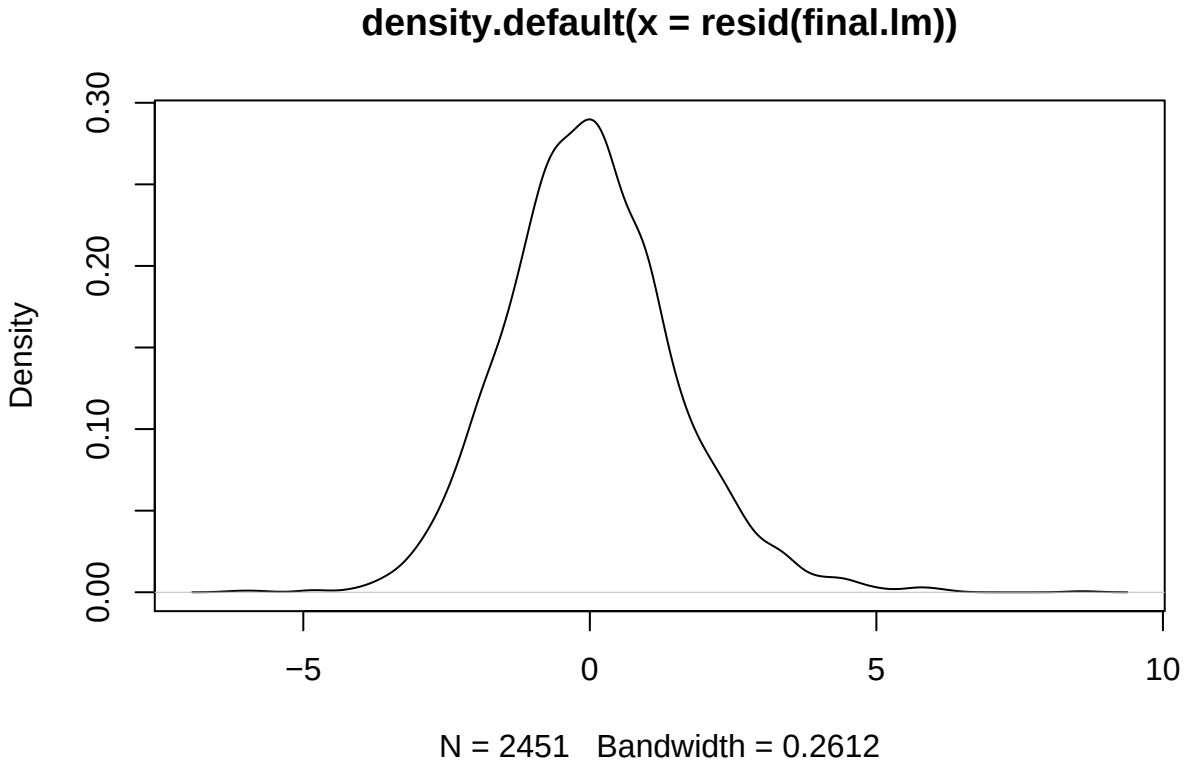

Residuals vs Fitted
Residuals
Fitted values
lm(prevalence ~ access * obese + assistance + fastfood + uninsured13 * inco ...
43
378
71


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(prevalence ~ access * obese + assistance + fastfood + uninsured13 * inco ...
43
83
71






```
plot(density(resid(final.lm)))
```



Results

The simple linear model testing the response variable, prevalence of type II diabetes, and the treatment, poverty, shows a statistically insignificant relationship and an adjusted R-squared value of approximately 0 (0.0005897). This signifies that poverty alone is not sufficient for indicating causation of diabetes type II.

In contrast, the simple linear models testing median household income, uninsured rates, access to grocery stores, and leisure time suggest a statistically significant relationship when individually tested against the response variable. Respectively, the corresponding adjusted R-squared values are as followed: 0.3247, 0.1882, 0.03372, and 0.5495. The linear regression model for prevalence of type II diabetes and leisure time spent physically inactive yielded the strongest correlation when treated alone.

A multiple linear regression model testing all 9 treatments against the response variable yielded an adjusted R-squared value of 0.6496, which is stronger than the max correlation from the simple regressions. Each treatment displayed a statistically significant relationship with the response variable, ($Pr(>|t|) < 3.68e\text{-}10$.

Because the multiple linear regression model produced a noticeable increase in the adjusted R squared, I was curious to test the importance of each variable and see which treatments produced the greatest percent mean square error. A second multiple linear regressions was constructed using the randomForest package. The result is quite significant and shows that the randomForest regression produced 500 trees that specified 74.13% of the variance could be explained. Based on the variance importance plot, the importance of each variable is ranked

as followed from most to least: leisure, assitance, insecurity, access, income, poverty, smokers, uninsured13, and fast food.

Two additional linear models were constructed to include interactions among variables. Those who are living in poverty or those who have low income may not have the time or resources to exercise and be active. Additionally, access to grocery stores is another important feature of living in poverty. People of lower incomes may not have the ability to purchase healthier foods; thus, hindering their personal nutrition. In order to adjust the linear models for these specific scenarios, interactions among the treatment variables were added.

The original interaction-designed multiple linear regressions model, named smoke.lm, featured three interactions: income with uninsured13, poverty with assitance, and poverty with leisure. The adjusted R-squared value for the model was 0.6631, which is just slightly stronger than what has been previously recorded. Each treatment again showed statistically significant relationships with the response variable except for income, poverty, and smokers. Hence, this outcome suggests the relationship between income, poverty, and smokers alone is not statistically significant.

The updated and final interaction-designed regression model, named final.lm, involved testing with three interactions; however, this time, the interaction between uninsured rate variable and income, the interaction between leisure time and smoking, and the interaction between access and obesity. The adjusted R-squared increased to 0.6671 and every single treatment yielded a statistically significant relationship with the response variable except for income, poverty and obese.

Plotting the different linear models and observing the residual plots helped indicate a distribution for the data. The quantile normal plots of the linear residuals from the multiple linear regression models appear to be approximately linear; thus, supporting the condition of linear regression that the error terms are normally distributed with mean, mu, and variance, sigma squared. This is also observed by constructing a density plot of the residuals. Each regression had approximately bell shaped density plots.

Conclusion

In conclusion, the regression analysis formed by the several multiple linear regression models show a relatively good relationship between the response variable and the treatments with apporximatley 70 percent of the variance explained. It would be unconvincing to decide that poverty alone is causing the higher prevalence in type II diabetes based on model correlations; however, one could say that the interaction between variables such as low income and additional features of poverty that incude: leisure time, health insurance, access to healthy foods, smoking, and food insecurity, show a stronger effect on causation of diabetes by poverty.