The Relationship Between Household Income and Overall Health Conditions in American Counties

Xueru Xie

University of California, San Diego

POLI 5, Professor Xuan

Summer Session 2, 2020

Abstract

This report analyzes the relationship between household income and overall health conditions in American counties in 2016. Specifically, this report takes a close look at the median household income in every county in America. To evaluate the overall health conditions in American counties, the report considers factors of adult smoking, adult obesity, alcohol drinking, sexually transmitted infections, and teen births. The question is, do higher household income counties tend to have better overall health conditions? The results illustrate that higher household income counties do have better health conditions in most of the factors than lower household income counties.

Introduction

Is America healthy? I believe this question has been asked many times by experts, public figures, and ordinary people. However, before answering the question, we need to ask ourselves, what is considered healthy? I think we can all agree that smoking is not a healthy habit. In 2016, the prevalence of current cigarette smoking among adults was 15.5%, which was a significant decline from 2005 (20.9%); however, no significant change has occurred since 2015 (15.1%) (Jamal A, Phillips E, Gentzke AS, et al, 2016). From this data, we can see that although the smoking rate has been largely decreased since 2005, it is still prevailing among many Americans. Another unhealthy habit when thinking about smoking would be drinking alcohol. We all know that alcoholism could cause damages. Alcohol-impaired-driving is a serious issue around the world. Of the 10,497 people who died in alcohol-impaired-driving crashes in 2016, there were 6,479 drivers (62%) who had BACs of .08 g/dL or higher, according to the Fatality Analysis Reporting System (Anonymous, 2020). There are other unhealthy habits such as having obesity, getting STDs, and teen births. Regardless, these five factors will determine America's healthiness in my report because they are less controversial. In order to measure America's healthiness, I will analyze the dataset by looking at counties. I believe that a county could reveal a lot of information about the people because different counties provide different quality of life for residents. Some counties could contain a population of certain racial groups, some counties could have higher productions in productions, and some counties could have poor infrastructures. More importantly, different counties contain citizens that have all kinds of health conditions. Thus, I believe that there is a relationship between counties and health conditions. Furthermore, since counties can have different people with different wealth, I also believe that counties with higher

household incomes will have better health outcomes because higher income neighborhoods have better living conditions, thus they have less sickness than poor neighborhoods.

Data

The dataset that I use is the County Health Rankings 2018, from the County Health Rankings website https://www.countyhealthrankings.org/. The dataset contains the measurements of all County Health Rankings measures. Each measure variable is a county in America. Not only does it have the measure value, it also has the numerator, denominator, and confidence interval of the measure value. Each row contains the county code as FIPS County Code, and state code as FIPS State Code. Next, the rest of the dataset contains columns of measurements. In total, there are 520 columns and 3200 rows. In each column of measure values, the numbers are mostly ratio variables, meaning that they represent a proportion of the population in a particular county. Some of the numbers reveal the number statistics per 100,000 population, some of the numbers reveal the proportion of the total population in the county, some of the numbers are actually number counts. Since there are 520 columns, not all the data are collected from one single source. There are numbers from the CDC center, the American Community Survey, the Bureau of Labor Statistics, and more. The years of data varies from 2010 to 2017, but mostly from 2016. As a result, the report will focus on data from 2016 to maintain consistency. The measurements can be divided into several categories. According to the County Health Rankings Codebook, the measure IDs are divided into factors like health outcomes, health behaviors, social and economic factors, social and economic environment, demographics, and so on. Since our topic will focus on household income and health conditions, the factors for our experiment will become only health outcomes, health behaviors, and social

and economic factors. The new dataframe after the modification will have 6 columns instead of 520 columns. In the new modified dataframe, all the columns will be renamed to specific descriptions. For example, the original column name “measure_9_value” will become “Adult smoking”, which tracks the statistics of adult smoking in a county. Here is the data dictionary for the new dataframe.

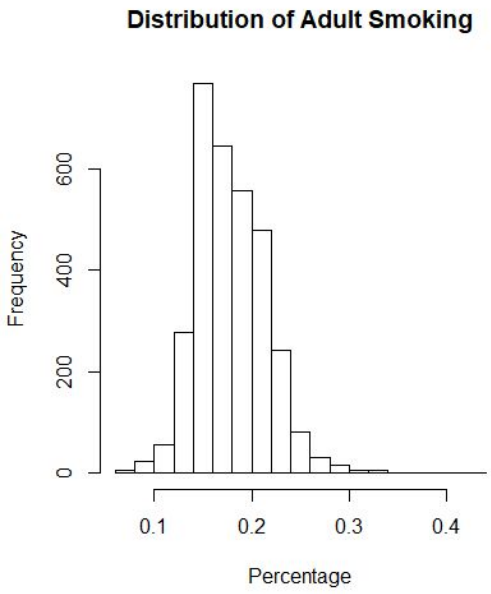
- Adult smoking
- Adult obesity
- Excessive drinking
- Sexually transmitted infections
- Teen births
- Median household income

Before I apply any methods to examine the relationship, I construct a summary table for each column below.

Adult smoking	Adult obesity	Excessive drinking
Min. :0.06735	Min. :0.1280	Min. :0.09265
1st Qu.:0.15234	1st Qu.:0.2880	1st Qu.:0.15102
Median :0.17316	Median :0.3170	Median :0.17403
Mean :0.17867	Mean :0.3143	Mean :0.17431
3rd Qu.:0.20275	3rd Qu.:0.3440	3rd Qu.:0.19678
Max. :0.42754	Max. :0.4780	Max. :0.29440
NA's :7	NA's :6	NA's :7

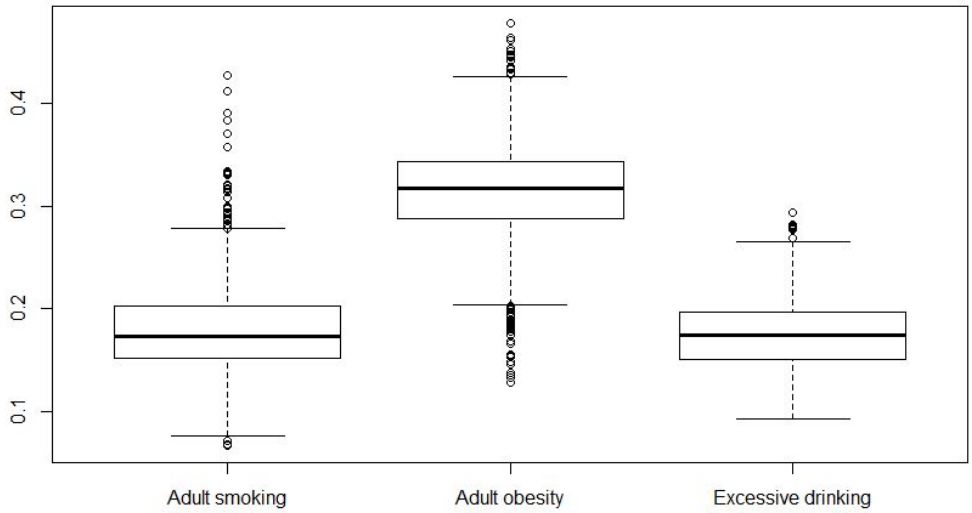
Sexually transmitted infections	Teen births	Median Household Income
Min. : 38.7	Min. : 2.822	Min. : 22045
1st Qu.: 208.7	1st Qu.: 22.402	1st Qu.: 41128
Median : 305.1	Median : 32.518	Median : 47737
Mean : 366.3	Mean : 34.292	Mean : 49663
3rd Qu.: 457.6	3rd Qu.: 44.270	3rd Qu.: 55452
Max. :2889.7	Max. :112.965	Max. :134609
NA's :173	NA's :134	NA's :7

This summary table above shows the minimum, 1st quartile, median, mean, 3rd quartile, maximum, and null values for each column. For the Adult smoking column, we could see that the smoking rate in a county can go as low as 6.3% and go as high as 42.8%. Next, we detect that the mean and median were quite similar. This makes me wonder whether the adult smoking rate is normally distributed.



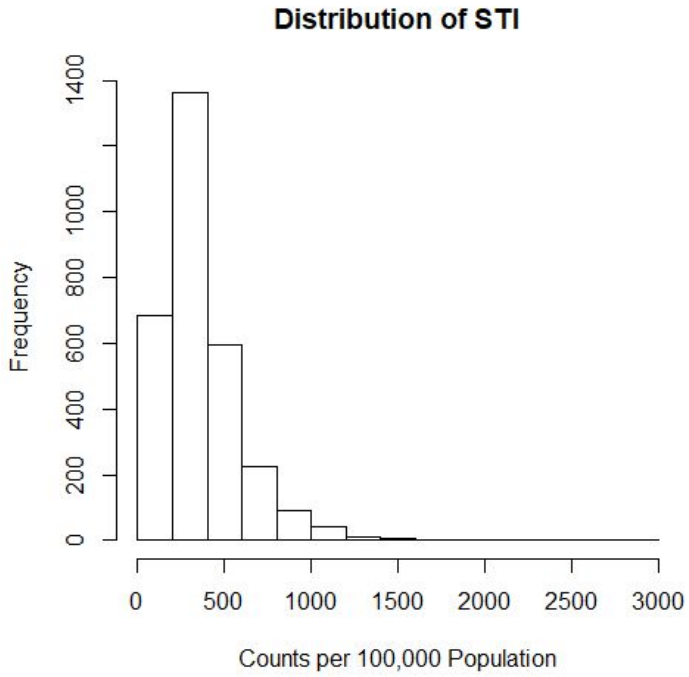
Based on this histogram, we can infer that the adult smoking rate is roughly a normal distribution, with a slight skew to the right. The impression would be that Americans are actually doing better in terms of adult smoking because most of the percentages lie below the median percentage, which means that many counties had fewer smokers than usual.

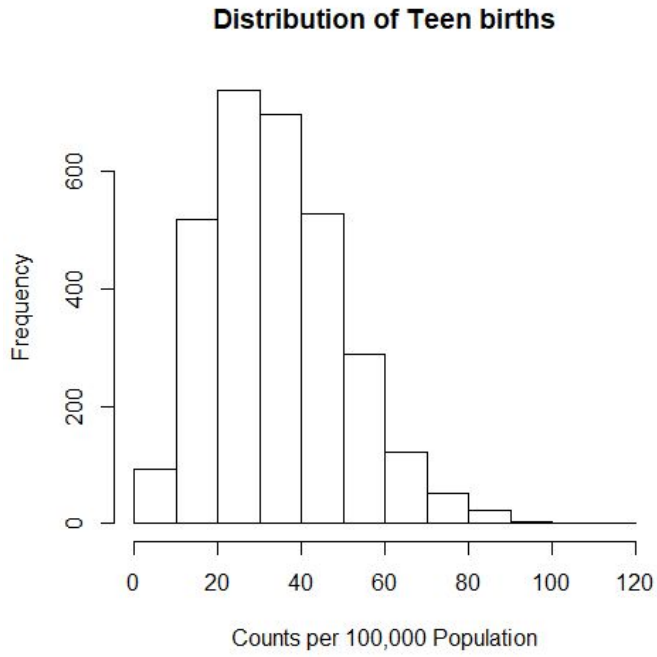
Next, I decide to plot box plots for columns Adult smoking, Adult obesity, and Excessive drinking in order to visualize the percentage distribution of these three health outcomes. I choose to neglect columns Sexually transmitted infections, and Teen births because the columns contain counts rather than proportions.



These boxplots are able to show the characteristics of each health outcome because they show the "severeness" of each lifestyle. We can see that adult obesity had the most severeness because the mean proportion is well above the other two. Which means that about ⅓ of American people have obesity, while only around ⅕ of American people have been smoking or excessive drinking.

Next, two other health outcomes are analyzed using histograms. These factors, sexually transmitted infections and teen births, are measured by the counts over 100,000 population.

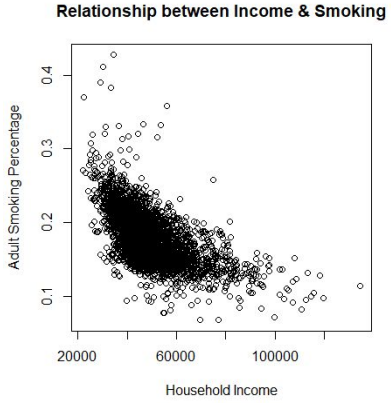

Distribution of STI
Frequency
0
200
400
600
800
1000
1400
0
500
1000
1500
2000
2500
3000
Counts per 100,000 Population

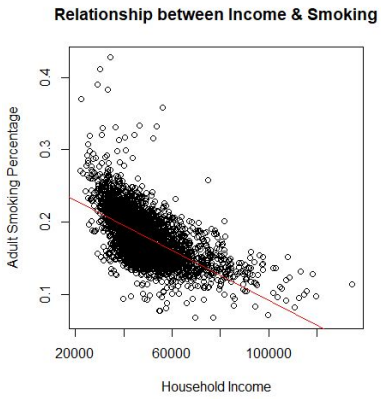

Distribution of Teen births
Frequency
0
200
400
600
0
20
40
60
80
100
120
Counts per 100,000 Population

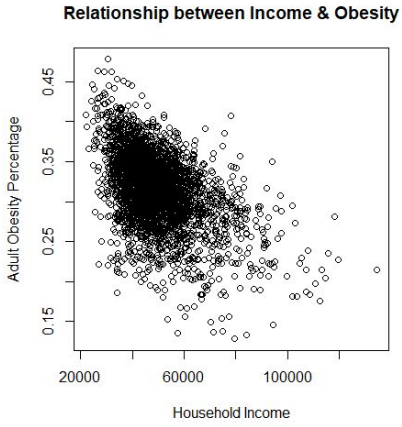
For column Sexually transmitted infections (STI), the histogram is skewed right. The mean in this case would be larger than the mean and the mode. For column Teen births, the distribution is roughly Gaussian, and the mean is around 34 cases per 100,000 population.

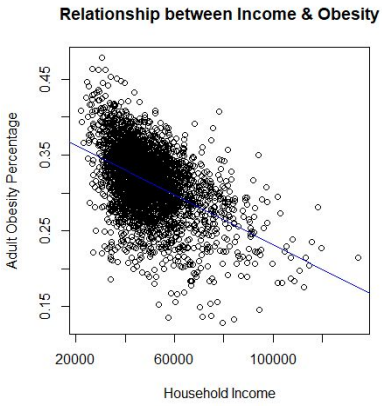
Methods

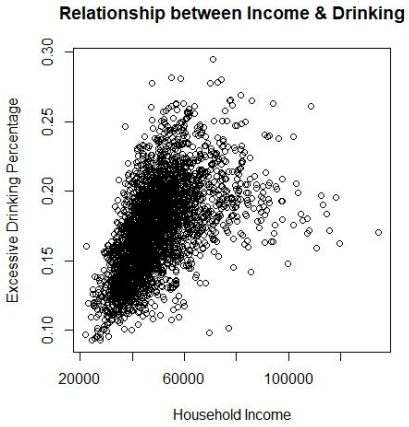
To examine the relationship between household income and health outcomes, it is necessary to build scatter plots around the independent and dependent variables. Doing so would demonstrate how the dependent variable would be related to the independent variable.Our independent variable is the median household income, and the dependent variables are the five health outcomes. Here are the scatter plots and regression lines included scatter plots below.

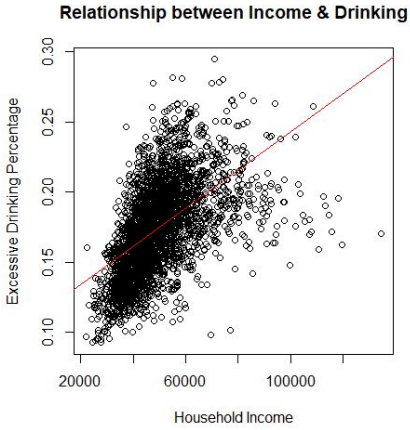


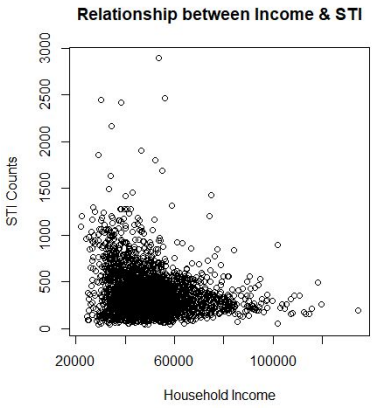


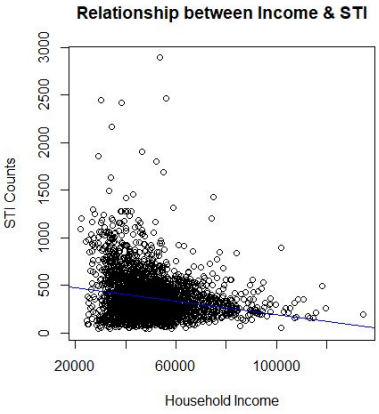


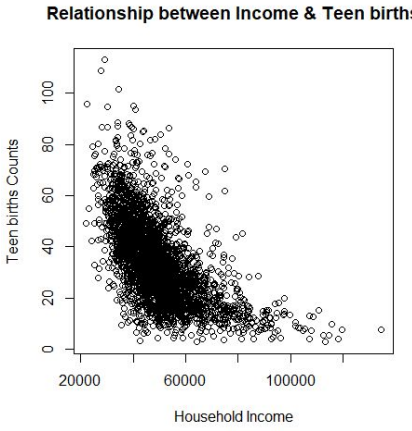


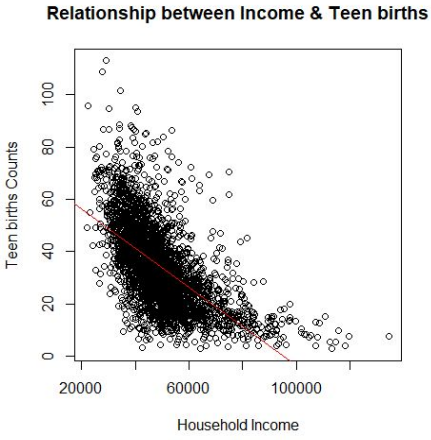

Relationship between Income & Drinking
Excessive Drinking Percentage
0.10
0.15
0.20
0.25
0.30
20000
60000
100000
Household Income


Relationship between Income & Drinking
Excessive Drinking Percentage
0.10
0.15
0.20
0.25
0.30
20000
60000
100000
Household Income


Relationship between Income & STI
STI Counts
0
500
1000
1500
2000
2500
3000
20000
60000
100000
Household Income


Relationship between Income & STI
STI Counts
0
500
1000
1500
2000
2500
3000
20000
60000
100000
Household Income


Relationship between Income & Teen births
Teen births Counts
0
20
40
60
80
100
20000
60000
100000
Household Income


Relationship between Income & Teen births
Teen births Counts
0
20
40
60
80
100
20000
60000
100000
Household Income

Based on the plots above, there are 3 health outcomes that have shown a negative relationship with the median household income. On the other hand, the relationship between median household income and excessive drinking seems to show a positive correlation. Lastly, although the relationship between median household income and STI seems to be negatively correlated, the line of best fit seems to have a small slope, meaning that the negative correlation is not significant. In order to see a detailed relationship between the independent and dependent variables, we need a linear model in this context.

Results

- Adult smoking

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.101990 -0.020582 -0.002144  0.018404  0.222407

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     2.644e-01  2.045e-03  129.28   <2e-16 ***
shrinked$`Median Household Income` -1.726e-06  3.987e-08  -43.31   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02902 on 3190 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.3703,    Adjusted R-squared:  0.3701
F-statistic:  1876 on 1 and 3190 DF,  p-value: < 2.2e-16
```

This regression table tells us that when the median household income increases by 1%, the percentage of adult smoking decreases by -1.726 x 10^-6 on average, all else equal. When the median household income is zero, the percentage of adult smoking is around 26%. The p-value is less than 2.2 x 10^-16, meaning that the result is very significant. The R-squared value is 0.37, that means 37% of the adult smoking percentage variation can be explained by the median household income.

- Adult obesity

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.16657 -0.02380  0.00345  0.02678  0.13930

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    3.957e-01  2.809e-03  140.85   <2e-16 ***
shrinked$`Median Household Income` -1.637e-06  5.475e-08  -29.91   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03986 on 3191 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.2189,    Adjusted R-squared:  0.2187
F-statistic: 894.4 on 1 and 3191 DF,  p-value: < 2.2e-16
```

This regression table tells us that when the median household income increases by 1%, the percentage of adult obesity decreases by -1.637 x 10^-6 on average, all else equal. When the median household income is zero, the percentage of adult obesity is around 40%. The p-value is less than 2.2 x 10^-16, meaning that the result is very significant. The R-squared value is 0.22, that means only 22% of the adult obesity percentage variation can be explained by the median household income.

- Excessive drinking

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.119557 -0.019430 -0.000611  0.018082  0.105928

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.066e-01  1.923e-03   55.43   <2e-16 ***
shrinked$`Median Household Income` 1.364e-06  3.748e-08   36.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02728 on 3190 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.2934,    Adjusted R-squared:  0.2932
F-statistic:  1324 on 1 and 3190 DF,  p-value: < 2.2e-16
```

This regression table tells us that when the median household income increases by 1%, the percentage of excessive drinking increases by 1.364 x 10^-6 on average, all else equal. When the median household income is zero, the percentage of excessive drinking is around 11%. The p-value is less than 2.2 x 10^-16, meaning that the result is very significant. The R-squared value is 0.29, that means 29% of the excessive drinking percentage variation can be explained by the median household income.

- Sexually transmitted infections

```
Residuals:
    Min      1Q  Median      3Q     Max
-400.83 -155.45  -52.52   99.78 2537.54

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.421e+02  1.684e+01    32.2   <2e-16 ***
shrinked$`Median Household Income` -3.538e-03  3.276e-04   -10.8   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 235.7 on 3025 degrees of freedom
  (173 observations deleted due to missingness)
Multiple R-squared:  0.03713,   Adjusted R-squared:  0.03681
F-statistic: 116.6 on 1 and 3025 DF,  p-value: < 2.2e-16
```

This regression table tells us that when the median household income increases by 1%, the number of STI decreases by 3.538 x 10^-3 on average, all else equal. When the median household income is zero, the count of STI is around 542. The p-value is less than 2.2 x 10^-16, meaning that the result is very significant. The R-squared value is 0.37, that means 37% of the variation of STI can be explained by the median household income.

- Teen births

```
Residuals:
    Min      1Q  Median      3Q     Max
-36.290  -8.344  -0.686   6.930  63.256

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     7.156e+01  9.005e-01   79.46   <2e-16 ***
shrinked$`Median Household Income` -7.505e-04  1.754e-05  -42.78   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.66 on 3064 degrees of freedom
  (134 observations deleted due to missingness)
Multiple R-squared:  0.3739,    Adjusted R-squared:  0.3737
F-statistic:  1830 on 1 and 3064 DF,  p-value: < 2.2e-16
```

This regression table tells us that when the median household income increases by 1%, the number of teen births decreases by 7.505 x 10^-4 on average, all else equal. When the median household income is zero, the count of teen births is around 72. The p-value is less than 2.2 x 10^-16, meaning that the result is very significant. The R-squared value is 0.37, that means 37% of the variation of teen births can be explained by the median household income.

Discussion

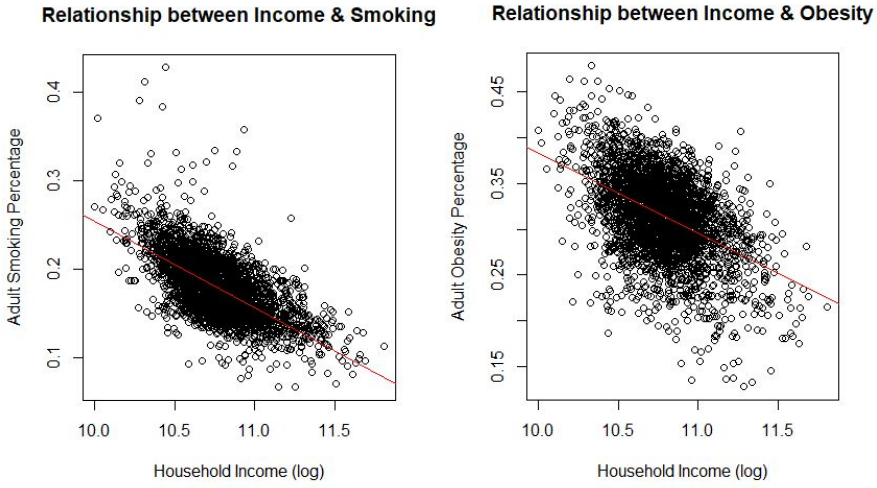
The linear regression model in R is able to answer my original question, that is what is the relationship between household income and health conditions? According to the results, there are significant correlations that show that counties with high median household income would have fewer percentages of adult smoking, adult obesity, and teen births. By looking at the scatter plots, we can see that there is a clear trend down in the percentages as the income increases. As a result, these three factors are able to prove my hypothesis. However, the result does not indicate any causality between income and the health conditions because there might be confounding variables. When we look at the percentage of excessive drinking, the result does not seem to agree with my assumption. The result shows that counties with higher income would have a

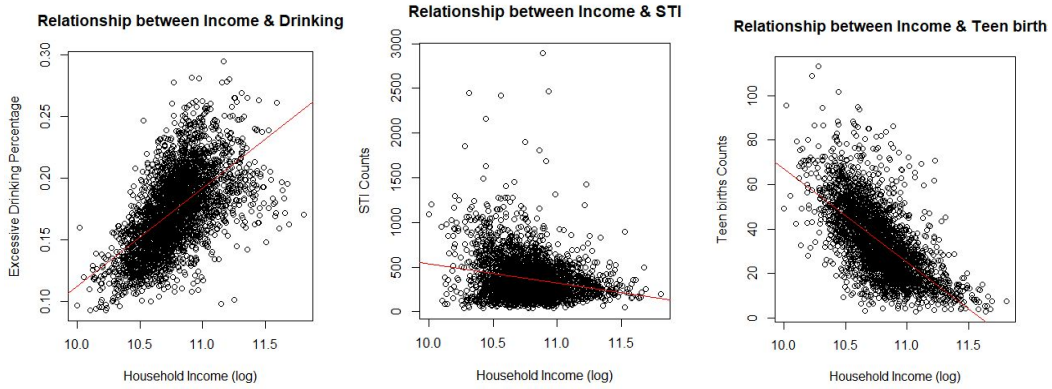
higher percentage of excessive drinkers. Based on this, I would then believe that people with higher income prefer drinking a lot because they have the time and money to do so. Lastly, when we look at the factor of STI, there is a weak negative correlation. That means although counties with higher incomes would have fewer STI, the differences between the number of people with STI are not significant between counties with various median household incomes.

One thing that stands out in the scatter plots is that there are many outliers for each graph. At first glance, most of the data are clustered together in a scatter plot, with some outliers around it. When looking at the outliers closely, we could see that most of the outliers are counties with very high median household incomes. That means that these rich counties sometimes do not fall into the formula in our linear regression models. I believe that this phenomenon makes sense because rich people have better living environments, thus they have the freedom to choose their health conditions based on their self-controls.

Robustness Check

When I look back at the scatter plots, I see that the units for x-axis and y-axis have very large differences. These differences cause very small coefficients in the linear regression models. So the purpose of this robustness check is to reshape the x-axis (median household income) in logarithmic form. That way, the scale between the independent and dependent variables will be much smaller, and the graph will have better shape.





	Dependent variable:				
	`Adult smoking` (1)	`Adult obesity` (2)	`Excessive drinking` (3)	`Sexually transmitted infections` (4)	`Teen births` (5)
`Median Household Income`)	-0.097***	-0.088***	0.080***	-213.506***	-41.951***
	(0.002)	(0.003)	(0.002)	(17.339)	(0.902)
Constant	1.228***	1.259***	-0.684***	2,668.438***	486.597***
	(0.022)	(0.031)	(0.021)	(187.011)	(9.733)
Observations	3,192	3,193	3,192	3,027	3,066
R2	0.417	0.222	0.354	0.048	0.414
Adjusted R2	0.417	0.222	0.354	0.047	0.413
Residual Std. Error	0.028 (df = 3190)	0.040 (df = 3191)	0.026 (df = 3190)	234.371 (df = 3025)	12.255 (df = 3064)
F Statistic	2,282.476*** (df = 1; 3190)	911.980*** (df = 1; 3191)	1,746.964*** (df = 1; 3190)	151.623*** (df = 1; 3025)	2,160.817*** (df = 1; 3064)

Based on these new graphs above, we could see that the shape of each graph fills up the box as a result of the logarithmic scaling. Although we rescale the x-axis, the results do not change too much. The correlations are still consistent with the original unscaled dataset. Looking

at the summary table, the coefficients now become bigger, which means they are easier to interpret. But again, the change in coefficients does not mean the results are different.

Conclusion

In conclusion, the results show that counties with higher median household income would have better health outcomes in some way. Counties with higher median household income tend to have lower percentages of adult smoking, adult obesity, and number of teen births. However, they would have a higher percentage of excessive drinking. In terms of sexually transmitted infections, there is not a strong positive or negative correlation that could describe the relationship with household income. Future research could use this report to further investigate the relationship between household income and health outcomes, such as finding causality and conducting observation studies.

References

Anonymous. (2020, May 14). Fatality Analysis Reporting System (FARS). Retrieved from https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars

Jamal A, Phillips E, Gentzke AS, et al. Current Cigarette Smoking Among Adults — United States, 2016. MMWR Morb Mortal Wkly Rep 2018;67:53–59.

Hales CM, Carroll MD, Fryar CD, Ogden CL. Prevalence of obesity among adults and youth: United States, 2015–2016. NCHS data brief, no 288. Hyattsville, MD: National Center for Health Statistics. 2017.