Spring 2022

Determinants of State Average Life Expectancy

Alex Grimes

Spencer Sprague

Cole Thomas

Alex Grimes, Spencer Sprague, and Cole Thomas

George Fox University

Determinants of State Average Life Expectancy

Faculty Advisor: Mitchell Priestley

Spring 2022

Abstract

In this paper, our dependent variable is average life expectancy by state in the United States. The purpose is to determine which factors have an impact on average life expectancy, as well as the magnitude of these impacts. While other studies have been conducted on life expectancy, our focus is different in a subtle but important way. Rather than centering on individual life expectancy, studying factors like "activity" or "genetics," we focus on the broader population in a state and which factors affect life expectancy on a macro-scale. There is little known about this topic that isn't in direct reference to life expectancy for a single person, and we found that some commonly held beliefs about life expectancy, which may be based on studies such as these, were contradicted by this study. For example, it is understood that women tend to live longer than men (Disabled World). Yet, in our analysis, although all but 4 states were majority-female, we found that states with higher proportions of men have higher average life expectancies. This appears to be contrary to the other findings, but we explore our hypothesis as to why this may be the case. The information contained in this study could be useful to state legislators for determining policies that affect the variables in question. For example, we found that obesity rates have a negative correlation with life expectancy. This understanding could be used to sponsor state health campaigns that attempt to reduce obesity rates, raising their state's average life expectancy. We used OLS regression analysis and found that the percentage of smokers had the largest impact on life expectancy in a state, while the least impactful, though still significant, factor we found was the percentage of uninsured individuals. The information provided here on these factors could, perhaps, embolden anti-smoking and health care campaigns in states across the country.

JEL classification: A1; I10; I13

Keywords: Life Expectancy; Percentage Regressors

Research Question

On a state-by-state basis, what factors impact average life expectancy, and what is the magnitude of their impact?

Section 1: Introduction

In 2021, the FDA advanced policy on banning certain types of cigarettes (U.S. Food and Drug Association, 2021). This policy is mainly focused on preventing young people from nicotine addiction, but the auxiliary effects of this policy will relate to average life expectancy, because cigarette smoking is a determinant of average life expectancy. The U.S. is still forming official opinions and policies on factors that impact life expectancy such as obesity and smoking. Understanding how factors like these affect average life expectancy is influential in policymaking, because it is crucial to understand, as well as possible, the items to which the policy relates.

Since we know that smoking decreases not only individual life expectancy but is a major factor in average life expectancy within a state's population, efforts to reduce smoking could be emboldened and become more successful. This is similarly true for obesity. Obesity is heavily linked with several of the leading causes of death in the United States, according to the Centers for Disease Control (No Date). While it is true that people rarely die directly from obesity, it is a major determinant of several leading causes of death, such as cardiovascular disease. Obesity affects such a large number of individuals in the U.S. that it has become a major determinant of average life expectancy (CDC, No Date), while other, similar factors (e.g., violent crime, traffic accidents, suicide, etc.) remain largely insignificant, as demonstrated in this study. Additionally, understanding the effects of factors like obesity and smoking may influence individual behavior.

A person who may be considering whether to start smoking (or someone who is considering quitting) may be inspired by the magnitude of the effect smoking has on a large population. This study will provide clarity on which factors have the greatest impact on average life expectancy and highlight them for future consideration.

There are only five states that have a majority male population. The other 45 states have a majority female population, albeit with varying margins. Surprisingly, we found that states with smaller margins had a higher life expectancy. That is, states whose male-female proportion was closer to 50-50. This is surprising because it is commonly understood that women tend to live longer than men, leading us to the hypothesis that the higher the population of women, the higher the life expectancy. However, this was not the case. The regressor with the largest coefficient, ***Mfrom50***, measures the "distance" between and "direction" of the male proportion and .50 (exactly half the population), then multiplies that number by 100 to get an integer. This produces a negative number for states whose male proportion is below .50 and a positive number for states with a male proportion over .50 (e.g., a .47 male proportion will produce the integer -3, and a .51 male proportion will produce the integer 1). ***Mfrom50*** has a positive relationship with a state's average life expectancy. This contradicts our hypothesis that a higher female proportion would produce a higher life expectancy. However, we must take care not to confuse the issue. This does not suggest that having "more males than females" in population would produce higher life expectancies; only that having a proportion of males that approaches .50 tends to increase life expectancy for that population. We do not have nearly enough observations (only five) of male-dominant populations to draw conclusions about them.

We used three other variables in our regression: ***Smokers***, ***Obesity***, and ***Uninsured***. ***Smokers*** and ***Obesity*** measured the percentage of smokers and self-reported obesity in each state,

and ***Uninsured*** measured the proportion of uninsured individuals in each state. Of these three, ***Smokers*** was the most impactful to life expectancy, followed by ***Obesity*** and ***Uninsured*** respectively. Smoking has a high impact on life expectancy which was what we expected, however, we were surprised to learn that it was more impactful than obesity. The idea that obesity is connected to the leading causes of death in the U.S. led us to the hypothesis that it would be the most impactful variable. That said, a weakness of our ***Obesity*** variable is that it is self-reported. The Centers for Disease Control (CDC) collected these data on a voluntary basis and did not provide any sort of BMI benchmark for reporting that we could identify. It's possible that, had they used a benchmark, the data would be different, because some people may incorrectly believe they are obese, while others may incorrectly believe they are not.

For ***Uninsured***, we expected it to be impactful and have a negative relationship with life expectancy, but we assumed that it would be only mildly impactful when compared to other factors. This proved to be the case with our regression, so our hypothesis was confirmed for this variable.

All in all, ***Mfrom50*** and ***Obesity*** surprised us by contradicting our hypotheses, but ***Smokers*** and ***Uninsured*** produced results that supported our hypotheses.

The paper first discusses our data overview which provides detailed information on every variable that we used. This includes those we measured but chose not to use in our regression analysis, as well as the reason why we chose not to use them. This will be followed by the methodology we employed to obtain our results, followed by our results and interpretations, and finally, the conclusions that can be drawn from our results.

Section 2: Data Overview

The data sets used in this study are sourced between the years 2018 and 2020 depending on the variable. When selecting our data sources, we attempted to collect data from the most recently available year, while still reliably sourcing our information. With this principle in place, all of the data we collected fell within a three-year period of each other. While not all data are sourced from the same year, this should not be a problem, as there is minimal variation year-to-year among the regressors we looked at. Seeing as all of the data sets are sourced between 2018 and 2020, we will be predominantly studying regressors based on data from before the COVID-19 global pandemic. Seeing as the global pandemic started in the late winter of 2020, portions of our data sets will include data occurring during the pandemic; however, the main purpose of this study is to analyze data independently from the global pandemic and not centered around the potential impact of the COVID-19 pandemic on our variables.

LifeExpectancy is our dependent variable. These data represent the average life expectancy at birth for each state in the year 2018. With these data, we found that the average life expectancy for all 50 states ranged between 74.4 years and 81.0 years, with a mean of 78.2 years (Life Expectancy at Birth by State). For all variables, see **Table 1** for the full summary statistics. The data set used is from the Center for Disease Control and Prevention (CDC) and is based on Census Data from the U.S. Census Bureau. Since we are collecting and analyzing data on a state-by-state basis in this study, much of our data is based on census data and survey data collected from the U.S. Census Bureau, which provides the CDC with census information. While these survey data are often self-reported, we attempted to collect data for each state using the most reliable and accurate sources possible. The CDC and the U.S. Census Bureau were some of the most trustworthy outlets to source our data.

Smokers, ***Obesity***, and ***SmallRiskFactors*** were sourced from the CDC. For our ***Smokers*** variable, we collected data from 2018 for the percentage of adults (18 or older) in a state who are currently smokers. The CDC described current smokers as "people who reported smoking at least 100 cigarettes during their lifetime and who, at the time they participated in a survey about this topic, reported smoking every day or some days" (Current Cigarette Smoking among Adults in the United States). For the ***Obesity*** regressor, we collected data on the percentage of people in 2020 who are obese for each state (Adult Obesity Prevalence Maps). Body Mass Index (BMI) is typically the metric used to determine if an individual is obese. Individuals with a BMI of over 30 are considered to be obese (What Is the Body Mass Index (BMI)). However, these data were based on individuals who self-reported being obese, which is a potential limitation to the validity of these data. Despite this, the CDC is a reliable outlet to source this information. Lastly, from the CDC, we collected data for our ***SmallRiskFactors*** variable. Within this variable, we included drug overdose deaths per 100,000, suicides per 100,000, and vehicle accident deaths per 100,000 for each state in 2020 (see Variable Descriptions below). Data for drug overdoses per 100,000 and suicides per 100,000 were collected via the CDC (Drug Overdose Mortality by State; Stats of the State – Suicide Mortality). However, data for vehicle accident deaths per 100,000 were collected from the Insurance Institute for Highway Safety, a nonprofit organization that sources data from the U.S. Department of Transportation's Fatality Analysis Report (Fatality Facts 2019: State by State).

Another source we used for data collection was the Kaiser Family Foundation (KFF), a non-profit organization that uses the Census Bureau's American Community Survey (ACS) to obtain data. We utilized KFF to find our data for the ***Uninsured*** variable. For this variable, we collected data on the proportion of the population that does not have health insurance for each

state in 2019 (Health Insurance Coverage of the Total Population). For the ***Mfrom50*** variable, we also obtained data from KFF, giving us the proportion of males and females for each state in 2019 (Population Distribution by Sex (CPS)).

To find high school graduation rates for our ***HighSchool*** regressor, we used data from the National Center for Education Statistics (NCES), an organization that collects education-related data from around the country. Using NCES, we found data for the average public high school graduation rate in each state for 2018 (*Public High School Graduation Rates*). One limitation of this regressor is the high school graduation rates only includes public high schools.

Another source we utilized was the Federal Bureau of Investigation (FBI). We gathered census data from the FBI to get observations for our ***ViolentCrimeRate*** variable. This data set measures the number of violent crimes per 100,000 people in each state for the year 2019 (Crime in the United States).

Lastly, for our ***MHI1k*** regressor, we collected data directly from the U.S. Census Bureau, which also relies upon surveys to collect their data. From the U.S. Census Bureau, we obtained data measuring the median household income for each state in the year 2019 (Bureau, US Census).

Table 1

Dropping MHIlk: string-valued series

	Mean	Median	Minimum	Maximum
LifeExpectancy	78.220	78.650	74.400	81.000
Mfrom50	-0.96800	-1.0000	-2.5000	0.90000
Smokers	16.648	16.250	9.0000	25.200
Uninsured	0.084680	0.079500	0.030000	0.18400
Obesity	32.214	32.050	24.200	39.700
ViolentCrimeRate	365.48	349.65	115.20	867.10
SmallRiskFactors	57.104	55.100	38.200	115.30
HighSchool	89.600	90.000	83.000	94.000

	Std. Dev.	C.V.	Skewness	Ex. kurtosis
LifeExpectancy	1.7167	0.021947	-0.56097	-0.56331
Mfrom50	0.76783	0.79322	0.41236	-0.21020
Smokers	3.3158	0.19917	0.26606	-0.11563
Uninsured	0.030655	0.36201	0.77480	0.79963
Obesity	3.9874	0.12378	-0.11685	-0.73936
ViolentCrimeRate	151.50	0.41452	1.2154	2.1577
SmallRiskFactors	14.462	0.25326	1.4537	3.5557
HighSchool	2.7255	0.030419	-0.46430	-0.68609

	5% perc.	95% perc.	IQ range	Missing obs.
LifeExpectancy	74.875	80.635	2.2750	0
Mfrom50	-2.1350	0.59000	1.1250	0
Smokers	11.640	23.015	4.6250	0
Uninsured	0.042100	0.14075	0.040000	0
Obesity	24.455	39.045	6.5750	0
ViolentCrimeRate	169.61	701.85	179.30	0
SmallRiskFactors	38.710	83.345	17.100	0
HighSchool	84.550	93.000	5.0000	0

Variable descriptions

Mfrom50 is a measure of the male percentage from 50%. It is calculated by subtracting .50 from the male proportion observation, then multiplying that number by 100 to get an integer. If the male proportion of a state is .48, then this data point will be -2 (.48 - .50 * 100). A male proportion over .50 will result in a positive integer, while a proportion less than .50 will result in a negative integer. ***Mfrom50*** is a continuous, linear variable.

Smokers is the percentage of the population in each state that reported smoking at least 100 cigarettes in their lifetime and smoke at least one cigarette every other day at the time of reporting. It does not imply the intensity of each smoker (i.e., how many cigarettes they smoke per day), only the percentage that meets the aforementioned criteria. ***Smokers*** is a linear, percentage variable.

Uninsured is a linear, proportional variable simply measuring the proportion of the population of each state that does not have health insurance.

Obesity is the percentage of the population of each state that is obese. Similar to the ***Smoker*** variable, ***Obesity*** does not imply intensity (i.e. how obese the average person is), only the percentage of the population that self-reported their obesity to the CDC. It is a linear, percentage variable.

MHI1k is a linear, continuous variable measuring the Median Household Income for each state in thousands.

Highschool is a linear, percentage variable measuring the percentage of residents within a state that acquired a high school diploma.

ViolentCrimeRate is the number of violent crimes per 100,000 residents every year. It is a linear, continuous variable.

SmallRiskFactors is a collection of other variables that impact death rates within states that we think could have some bearing on this regression. We included drug overdoses, suicides, and vehicle accidents per 100,000 residents for each state annually and added them together. The purpose for this was mainly to ensure our model did not suffer from omitted variable bias, which it does. ***SmallRiskFactors*** is a linear, continuous variable.

Assumptions Evaluation

Assumption 1: Linearity

All our final variables (see regression 5 in Table 2) do have a linear relationship with our dependent variable, based on the individual scatter plots we created to measure the relationship. While our ***Mfrom50*** variable was transformed from male proportion observations, our other variables in our final regression remain as discovered. The linearity of these variables is supported by our R^2 being sufficiently high at .866.

Assumption 2 - Homoscedasticity

We ran residual plots for each of our final variables and identified heteroscedastic spread, so we coped with this by consistently employing heteroscedasticity-robust standard errors in order to ensure valid conclusions concerning the statistical significance of our coefficient estimates.

Assumption 3 - Independent Error Terms

Autocorrelation is not an issue with our data as it is not time-series bound. Since the data are not time-series bound, each succeeding residual is not affected by the preceding residual closest to it on the x-axis. There is no natural order to our regression error terms, and this is supported by the QQ plots in Appendix A.

Assumption 4 - Normal Errors

Regressions that lack normally distributed errors do not usually have any issues with reliability due to the central limit theorem, assuming those regressions have a large number of observations. Unfortunately, our regression is limited by the population of our data: 50 states. Due to this reason, we ran a QQ plot (see Appendix A) for each variable to check the normality of their distribution and found that they were indeed normally distributed. Therefore, the normal errors assumption is satisfied.

Assumption 5 - No Perfect Multicollinearity

Using a correlation matrix analysis and reasoning, we've determined that none of the regressors within our final model are perfectly multicollinear. Of the four final regressors that our model utilizes, each one represents a significantly distinct and unique characteristic in encapsulating an accurate model for average life expectancy by state.

Assumption 6 - No Omitted Variable Bias

With our R^2 being high as .866, our final regression model suffices for meeting this assumption. However, we recognize that there is a plethora of minor variables that influence life expectancy, and especially when they are added together, could become statistically significant. Our ***SmallRiskFactors*** variable was created as a control variable to account for these, but it was ultimately abandoned in the model due to its lack of economic significance. This is further explained below, in the Methodology section. Consequently, we have included some of the most statistically significant regressors within our linear regression model.

Section 3: Methodology

Regression Model 1:

$$LifeExpectancy = \beta_0 + \beta_1(Mfrom50) + \beta_2(Smokers) + \beta_3(Uninsured) + \beta_4(Obesity) + \beta_5(MHI1k) + \beta_6(Highschool) + \beta_7(ViolentCrimeRate) + \beta_8(SmallRiskFactors) + u$$

We began our methodology with Regression Model 1. We found that ***Mfrom50*** has a positive and constant relationship with life expectancy. This is contrary to what we expected going into the study, as women typically live longer than men on average. In fact, women outlive men on average in almost every society. In developed countries worldwide, the average life expectancy for women is 79 years old, while it is only 72 years old for men (Around the Globe, Women Outlive Men). Despite this notion, our research suggests that the higher proportion of men a state has, the higher the average life expectancy will be for that state. We don't know exactly why this is, but the hypothesis that made the most sense to us was that males tend to go into professions that have a strong impact on the average life expectancy for their state's population. We thought of professions like doctors, engineers, police, etc. These professions, which are male dominated, likely have a powerful, positive influence over the life expectancy of the entire population. So, a state with a higher proportion of males will likely have a higher proportion of doctors, engineers, and police officers than other states with a lower proportion of males, and this impact more than outweighs the slight difference in life expectancy between males and females.

Smokers has a negative and constant relationship with life expectancy. This is likely because smoking is highly correlated with other health conditions that dramatically lower an individual's life expectancy, such as cancer and heart disease (Health Effects). Furthermore,

when a high percentage of a population smokes, the life expectancy for that population decreases quite significantly.

Obesity has a negative and constant relationship with life expectancy. This is likely because obesity is linked with many of the leading causes of death across the nation. A state with more people reporting obesity will likely have higher percentages of the health conditions linked with obesity as well, which would drastically decrease life expectancy. We are only surprised that the coefficient of this variable wasn't larger than it already is.

Uninsured has a negative and constant relationship with life expectancy. This means that a higher proportion of uninsured individuals in a state is correlated with a lower life expectancy in that state. We hypothesize that this is because uninsured individuals would be much more hesitant to seek medical and preventative care when they need it, due to financial reasons. There may also be other, more subtle causes for this, but this hypothesis seems to provide the most clarity for this relationship.

MHI1k and ***HighSchool*** were not statistically significant within any of our regressions, therefore their coefficients are unreliable, and we did not use them beyond our second regression.

For our final regression model, we will also not be including ***ViolentCrimeRate*** or ***SmallRiskFactors***. Despite ***ViolentCrimeRate*** being statistically significant at the 10% significance level and ***SmallRiskFactors*** being statistically significant at the 1% significance level in each of the regressions they were included in, these variables do not provide practical significance. This is due to the coefficients of our ***ViolentCrimeRate*** being too small. For example, in regression 3, ***ViolentCrimeRate*** has a coefficient of -0.0009. Consequently, a one unit increase in violent crimes per 100,000 would only lead to a 0.0009 year decrease in a state's average life expectancy. For ***SmallRiskFactors***, the main reason for removing it from our final

regression model is due to the fact that it is a conglomerate of three different factors. With this, we are unable to identify what an increase or decrease in this variable would consist of; thus, the practical significance of this variable is limited. However, as was explained above, in the Assumption Evaluation section, this variable could have potentially been used as a control to mitigate the effect of Omitted Variable Bias. This has resulted in the following final regression model.

Regression Model 5 (with coefficients):

$$n = 50,\ R^2 = 0.866$$

$$\widehat{LifeExpectancy} = \underset{(0.766)}{87.91} + \underset{(0.148)}{0.59}(Mfrom50) - \underset{(0.059)}{0.31}(Smokers) - \underset{(0.035)}{0.13}(Uninsured) - \underset{(0.043)}{0.09}(Obesity)$$

Technique (OLS Estimation):

We have utilized the OLS estimation (Ordinary Least Squares) technique. This is a worthwhile choice for an estimation technique because, as illustrated in the Assumptions Evaluation section, all the assumptions are met for a standard regression model. In other words, it was the best technique available, and our data satisfied the requirements to obtain reliable results using the technique. When using the OLS estimator technique, the regression model is assumed to have the following characteristics: linearity, constant error variance (homoscedasticity), independent error terms (see Appendix A for QQ plots), a normal distribution of the error's vertical spread, no multicollinearity, and exogeneity. The data we used in our model either satisfied each of these assumptions or we applied other techniques to cope with its lack of satisfaction, such as heteroscedasticity-robust standard errors, which provide more reliable results for data that are not homoscedastic.

Section 4: Results & Interpretations

With Regression Model 5, and all regressions throughout this study, ***Mfrom50***, ***Smokers***, and ***Uninsured*** are statistically significant at the 1% level. ***Obesity*** is significant at the 10% level using our final regression model, but in previous models it was significant even at the 1% level.

There is economic significance for the regressors ***Smokers***, ***Uninsured***, and ***Obesity***. Since all of these regressors are negatively correlated with a state's average life expectancy, there are real-world implications that present the need for policymaking decisions to be introduced. A recent study of 825 Norwegians aged 60 and older stated that their average 'preferred life expectancy' (PLE) was 91.4 years of age (Bowens). However, there is an interesting twist to this study. When participants were faced with hypothetical scenarios of serious health situations (e.g., dementia, chronic pain) with progressive aging, their PLE decreased. Subconsciously, this causes people who select a high PLE to self-examine their lifestyle habits and identify if their current course of action is going to lead them to a long, happy, healthy life. Though the United States is not perfectly akin to Norway in terms of cultural beliefs and attitudes, there is still a consensus that a long life can be a happy one. With this study in mind, there might be economic significance for Americans to conduct a self-examination of their lifestyles to identify what might prevent them from living a long, happy life. For example, an American who is a current smoker that has the same desire to live past their 80s without risk of cancer/heart disease will likely consider the idea of quitting smoking. The same logic applies to Americans who are obese and those who are uninsured. However, ***Mfrom50*** does not hold economic significance, because it would be nonsensical to advocate increasing the male proportion of a state's population. We believe the economic significance of the following regression results, other than ***Mfrom50***, illustrated in Table 2 below, is apparent.

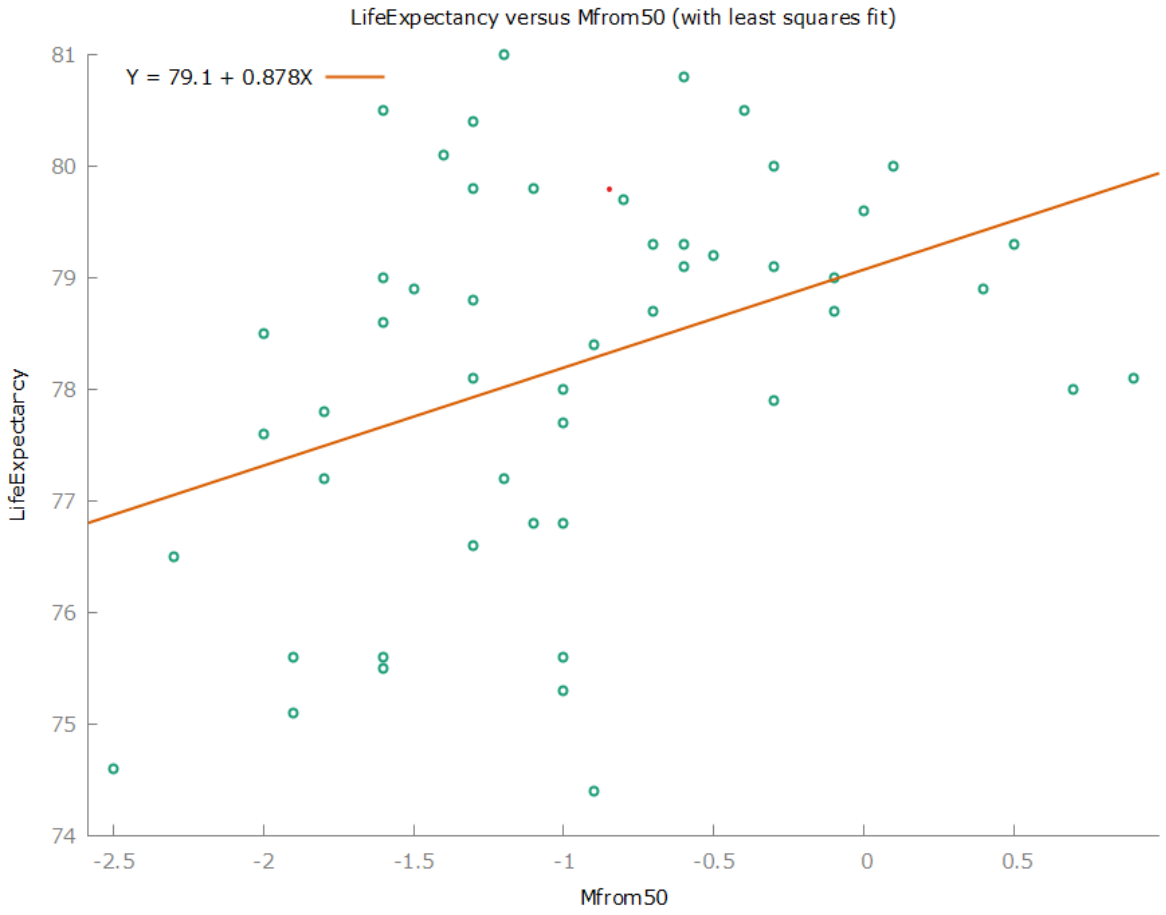
Table 2

Results of Average Life Expectancy by state on proportional variables gender, smokers, uninsured, obesity, and highschool, and other vairables median household income, violent crime rate, and other small risk factors related to life expectancy, using data from a myriad of sources, including the CDC and Census Bureau.

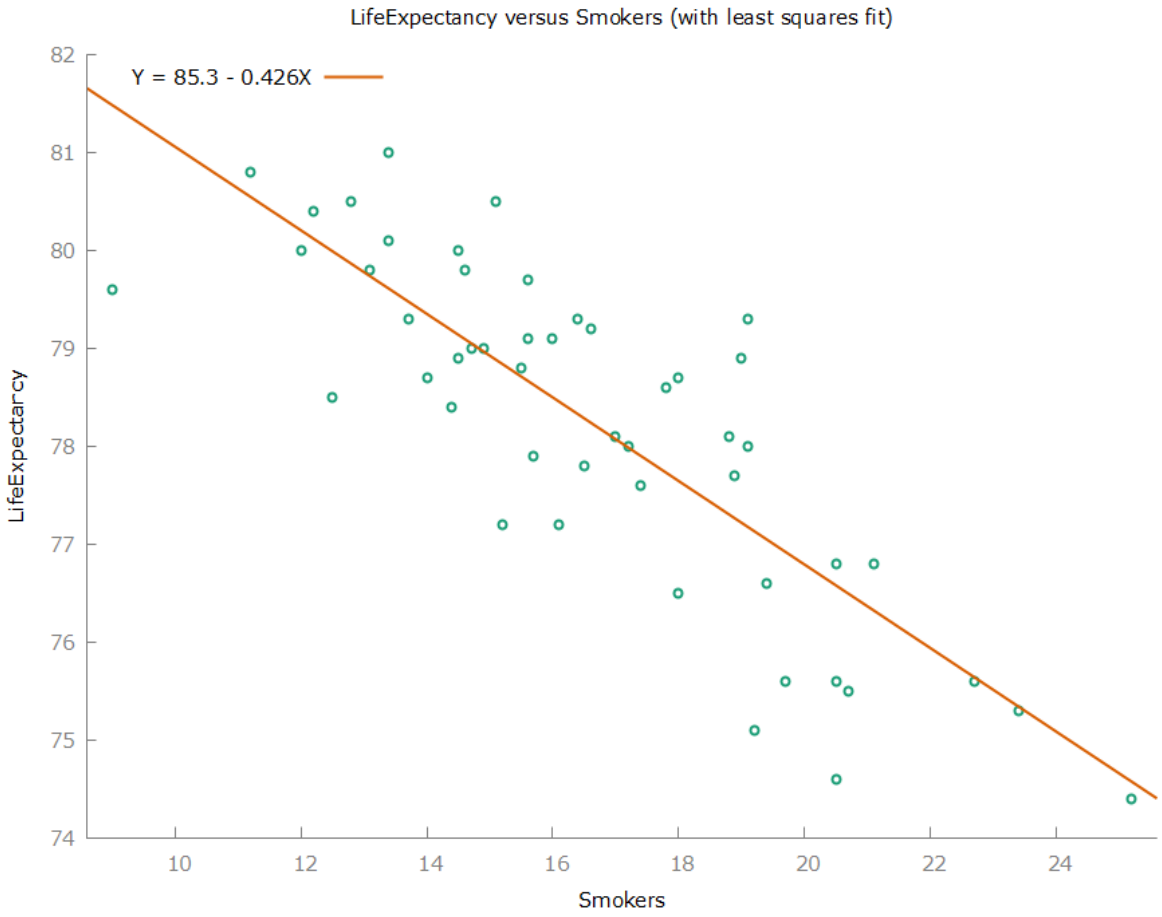
Dependent variable: average life expectancy by state.					
Regressor	(1)	(2)	(3)	(4)	(5)
Mfrom50 (X_1)	0.62 ***	0.52 ***	0.48 ***	0.48 ***	0.59 ***
	(0.153)	(0.133)	(0.124)	(0.121)	(0.146)
Smokers (X_2)	-0.31 ***	-0.21 ***	-0.21 ***	-0.22 ***	-0.31 ***
	(0.059)	(0.057)	(0.058)	(0.056)	(0.059)
Uninsured (X_3)	-0.14 ***	-0.12 ***	-0.12 ***	-0.13 ***	-0.13 ***
	(0.034)	(0.032)	(0.033)	(0.031)	(0.035)
Obesity (X_4)	-0.09 **	-0.11 ***	-0.11 ***	-0.11 ***	-0.09 *
	(0.043)	(0.037)	(0.038)	(0.039)	(0.043)
MHI1k (X_5)	-0.005	-0.005			
	(0.007)	(0.006)			
Highschool (X_6)	0.0002	0.0002			
	(0.0004)	(0.0003)			
ViolentCrimeRate (X_7)		-0.001 *	-0.0009 *		
		(0.0006)	(0.0005)		
SmallRiskFactors (X_8)		-0.03 ***	-0.03 ***	-0.03 ***	
		(0.007)	(0.006)	(0.006)	
Intercept	88.02 ***	88.82 ***	88.73 ***	88.67 ***	87.91 ***
	(0.765)	(0.640)	(0.645)	(0.655)	(0.766)
Summary Statistics					
SER	0.67	0.54	0.53	0.54	0.65
R^2	0.868	0.918	0.916	0.910	0.866
Adj. R^2	0.850	0.902	0.904	0.900	0.855
n	50	50	50	50	50

Results by Variable:

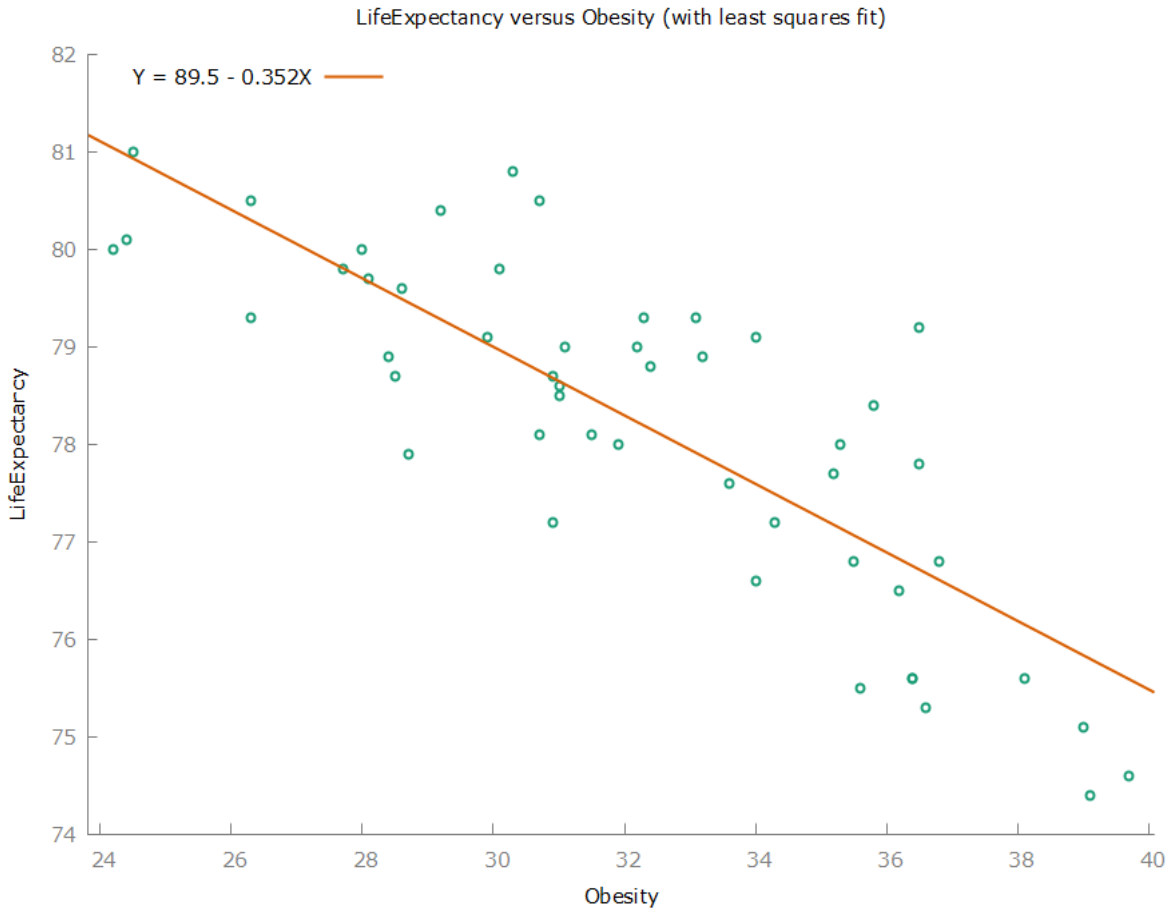
Mfrom50 is calculated from proportional data and is correlated with a .59-year increase in a state's average life expectancy. Ceteris paribus, this means that a one percentage point increase of the proportion of males in a state will result in an increase of .59 years, on average, of the average life expectancy of a given state. However, there are only 5 states where the male proportion is over .50. For this reason, we cannot derive any meaningful conclusion about states with a majority male population. This regression only provides insight on states whose male proportion is less than but approaching .50.



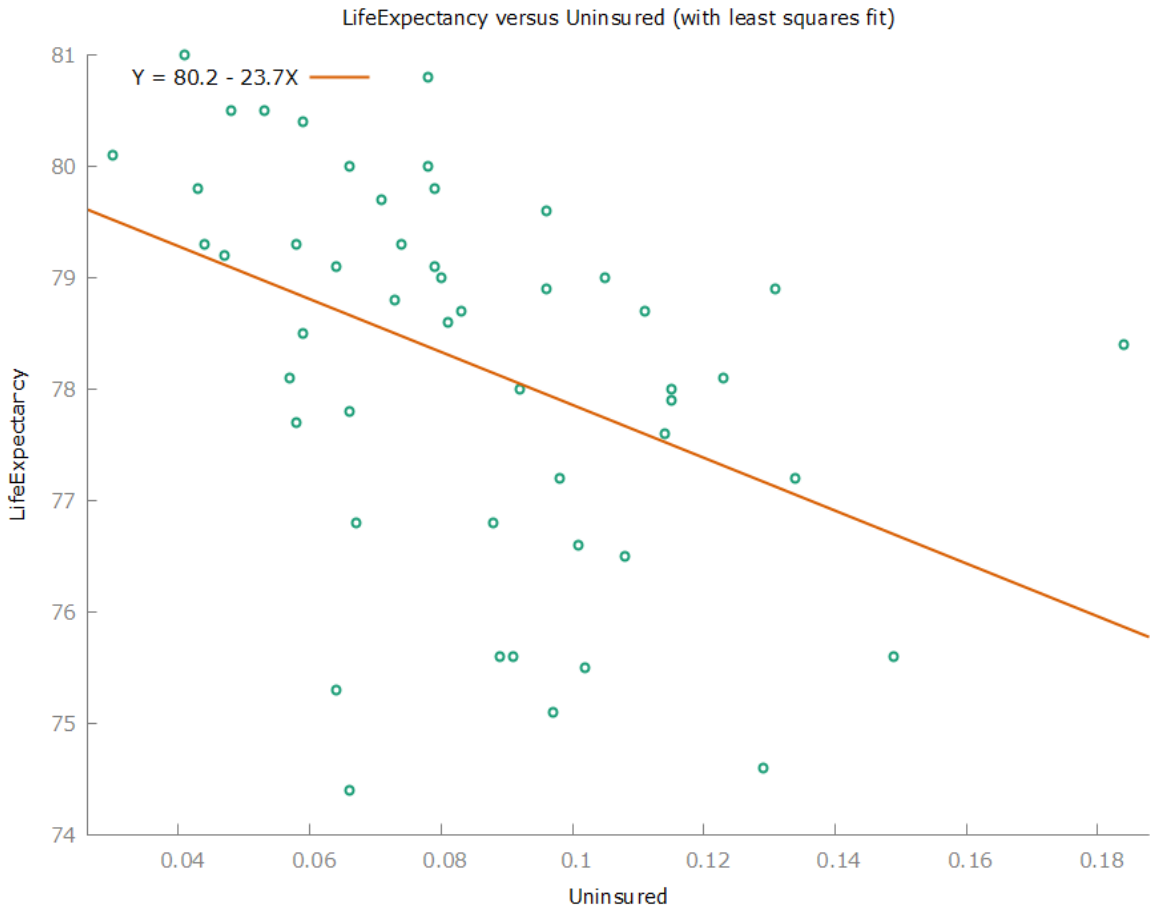
Smokers is comprised of percentage data and is correlated with a .31 year decrease in a state's average life expectancy. This means that a 3-percentage point increase in the percentage of smokers in a state will result in just under a 1 (.93) year decrease in a state's life expectancy. While this is useful for the purposes of this regression analysis, it ignores the intensity that one smokes. All smokers are counted the same regardless of whether they smoke one cigarette every other day or 2 packs of cigarettes per day. This will certainly leave out any nuance to be found within the data on this variable. For example, we don't know how impactful smoking 1 cigarette every other day is on an individual person's life expectancy, and we can't derive that answer with the data we have.



Obesity is also comprised of percentage data and is correlated with a .086 year decrease in a state's average life expectancy for every one percent increase in the obesity rate of a state. One of its greatest limitations is like that of ***Smokers*** in the sense that it ignores the magnitude of a state's obesity. One of the limitations of this CDC-based data is that it is self-reported. The CDC collected data by requesting that individuals report their obesity status, which is entirely subjective depending on the individual's concept of 'obesity.' This could cause the data to be largely unreliable.



Uninsured is also comprised of proportional data and is correlated with a .136 year decrease in a state's average life expectancy for every one percent decrease in a state's health insurance rate. One of the largest limitations of this variable is that it is specifically representing the lack of *health* insurance, not other forms of insurance. This is a limitation because while the regressor could identify the relationship between a state's health insurance status and average life expectancy, it does not consider other important forms of insurance that are prevalent amongst Americans.



Since our economically significant regressors all have a negative correlation with a state's average life expectancy, policymakers of state and local governments and municipalities can advocate for policy changes that assist Americans to increase their life expectancy by reducing the prevalence of these variables within their state. For example, the notion that smoking lessens average life expectancy is a common knowledge among adults but possibly not so common among children. Policymakers with the goal of lessening the rate of smoking in their state can advocate for heightened education on the effects of smoking in public school settings. Furthermore, since government taxation can act as a catalyst to incentivize behavior, there could be a state (or even federal) cigarette tax introduced to increase the price of those goods, making them less desirable for individuals with less marginal propensity to consume. Regarding obesity, further research can be paid for by the government in discerning what factors are causal to obesity as well as what combats it. This research can then be introduced into public school curricula to educate the youth about the causes and preventative measures regarding obesity. As for those who are uninsured for health insurance, further research will need to be sponsored by the government or other private sector entities in analyzing the demographics of these individuals. Rather than simply mandating that every American adult has health insurance, a more useful approach might be to survey individuals as to why they are lacking health insurance, as well as what can be done to fix that.

Section 5: Conclusion

We used the OLS estimator technique and found that the proportion of males in a population has a positive relationship with a state's average life expectancy. The percentage of smokers in a state also had a large impact on average life expectancy. The percentage of obese and proportion of uninsured individuals, however, had a lower impact on life expectancy than we anticipated.

We believe that attention should be given to policies regarding cigarette smoking, health insurance policies, and obesity. All these factors have a negative relationship with average life expectancy; ergo, to improve average life expectancy, decreasing the prevalence of smokers, people who are uninsured, and those that are obese will significantly improve a state's average life expectancy.

Although we understand that women tend to have a longer life expectancy than men, states with a male proportion approaching 50% will have a higher average life expectancy than states with a larger margin of female dominance in the population.

One of the greatest limitations of this research is omitted variable bias. While the four major regressors we utilized do a fantastic job at showcasing their relevance on average life expectancy, they represent an extremely small portion of the possible thousands of regressors that exist in affecting average life expectancy. We understand that the use of ***SmallRiskFactors*** as a control variable would have mitigated the effect of Omitted Variable Bias, however, there was no sensical way to explain the effect of the variable on a state's average life expectancy. It's also true that an all-encompassing model that utilizes all possible regressors would be unrealistic and unnecessary for this study. Results from a few regressors that have statistical and economic significance, combined with a sufficiently high R^2, are reliable.

Section 6: References

"Adult Obesity Prevalence Maps." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 27 Sep. 2021, https://www.cdc.gov/obesity/data/prevalence-maps.html.

"Around the Globe, Women Outlive Men." *PRB*, https://www.prb.org/resources/around-the-globe-women-outlive-men/.

Bowens, et al. "How Long Do People Want to Live?" Schizophrenia.com, 31 Aug. 2021, https://forum.schizophrenia.com/t/https-www-usnews-com-news-health-news-articles-2021-08-12-blood-test-spots-biological-markers-for-schizophrenia/247811.

Bureau, US Census. "2019 Median Household Income in the United States." *Census.gov*, 8 Oct. 2021, https://www.census.gov/library/visualizations/interactive/2019-median-household-income.html.

Centers for Disease Control. "Adult Obesity Causes & Consequences." No Date, https://www.cdc.gov/obesity/adult/causes.html

"Crime in the United States." *FBI*, FBI, 29 Aug. 2019, https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/topic-pages/tables/table-5

"Current Cigarette Smoking among Adults in the United States." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 17 Mar. 2022, https://www.cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/index.htm.

Disabled World. Longevity: Extending Life Span Expectancy. *Disabled World*, 5 Mar. 2022, www.disabled-world.com/fitness/longevity/

"Drug Overdose Mortality by State." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 1 Mar. 2022, https://www.cdc.gov/nchs/pressroom/sosmap/drug_poisoning_mortality/drug_poisoning.htm.

"Fatality Facts 2019: State by State." *IIHS*, https://www.iihs.org/topics/fatality-statistics/detail/state-by-state.

"Health Effects." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 28 Apr. 2020, https://www.cdc.gov/tobacco/basic_information/health_effects/index.htm.

"Health Insurance Coverage of the Total Population." *KFF*, 15 Nov. 2021, https://www.kff.org/other/state-indicator/total-population/?currentTimeframe=0&sortModel=%7B%22colId%22%3A%22Location%22%2C%22sort%22%3A%22asc%22%7D.

"Life Expectancy at Birth by State." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 17 Mar. 2022, https://www.cdc.gov/nchs/pressroom/sosmap/life_expectancy/life_expectancy.htm.

"Population Distribution by Sex (CPS)." *KFF*, 28 Sept. 2021, https://www.kff.org/other/state-indicator/population-distribution-by-sex-

cps/?currentTimeframe=0&sortModel=%7B%22colId%22%3A%22Location%22%2C%22sort%22%3A%22asc%22%7D.

"Public High School Graduation Rates." *National Center for Education Statistics*, https://nces.ed.gov/programs/coe/pdf/coe_coi.pdf.

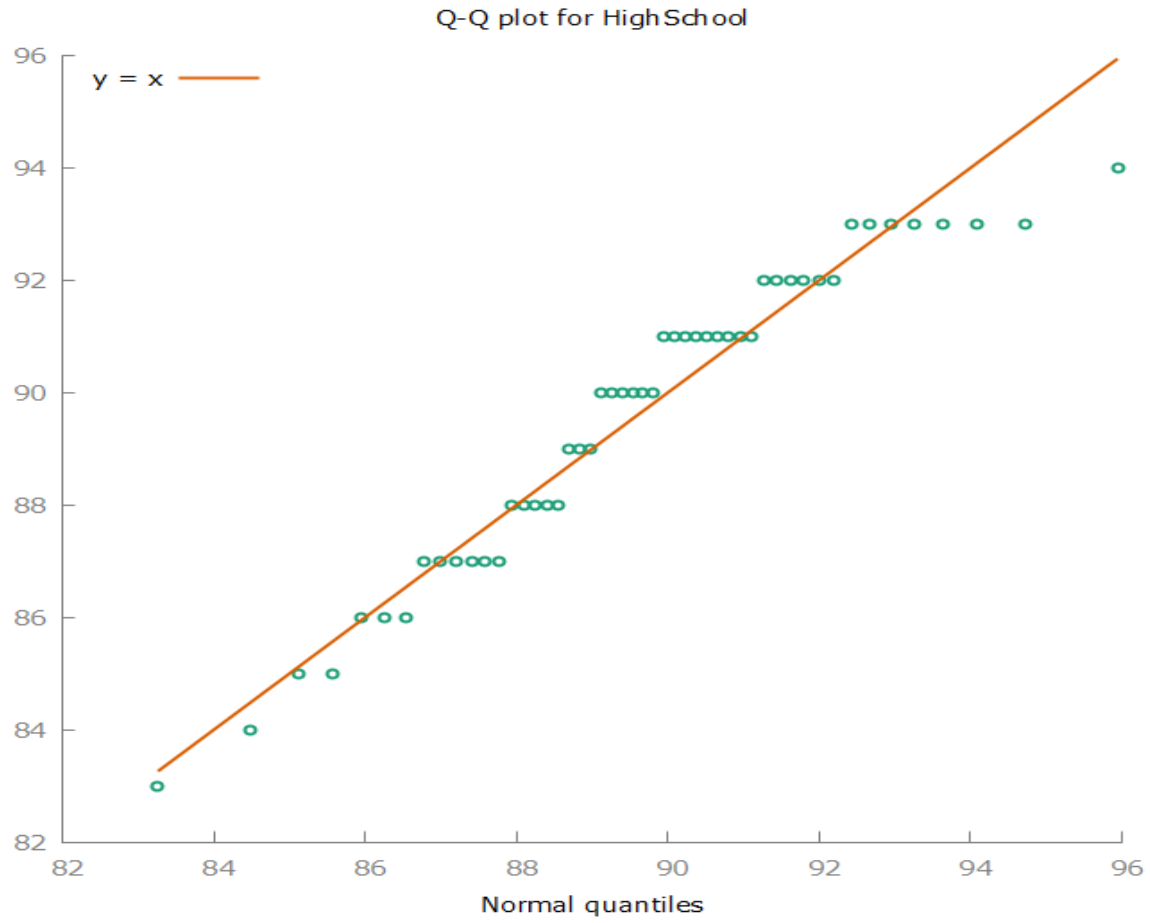
"Stats of the State - Suicide Mortality." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 1 Mar. 2022, https://www.cdc.gov/nchs/pressroom/sosmap/suicide-mortality/suicide.htm.

"FDA Commits to Evidence-Based Actions Aimed at Saving Lives and Preventing Future Generations of Smokers" [Press release]. *U.S. Food and Drug Administration*, 29 Apr. 2021 https://www.fda.gov/news-events/press-announcements/fda-commits-evidence-based-actions-aimed-saving-lives-and-preventing-future-generations-smokers

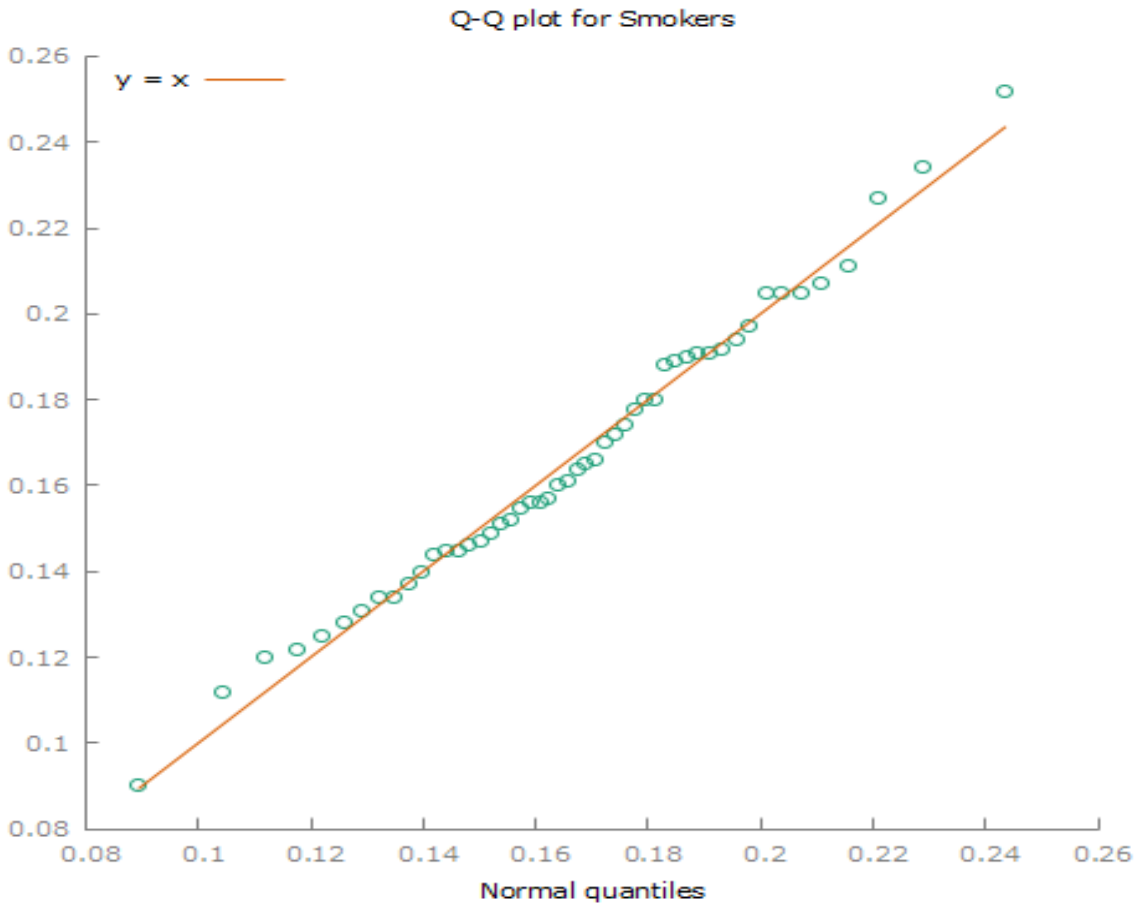
"What Is the Body Mass Index (BMI)?" *NHS Choices*, NHS, https://www.nhs.uk/common-health-questions/lifestyle/what-is-the-body-mass-index-bmi/.

Appendix A

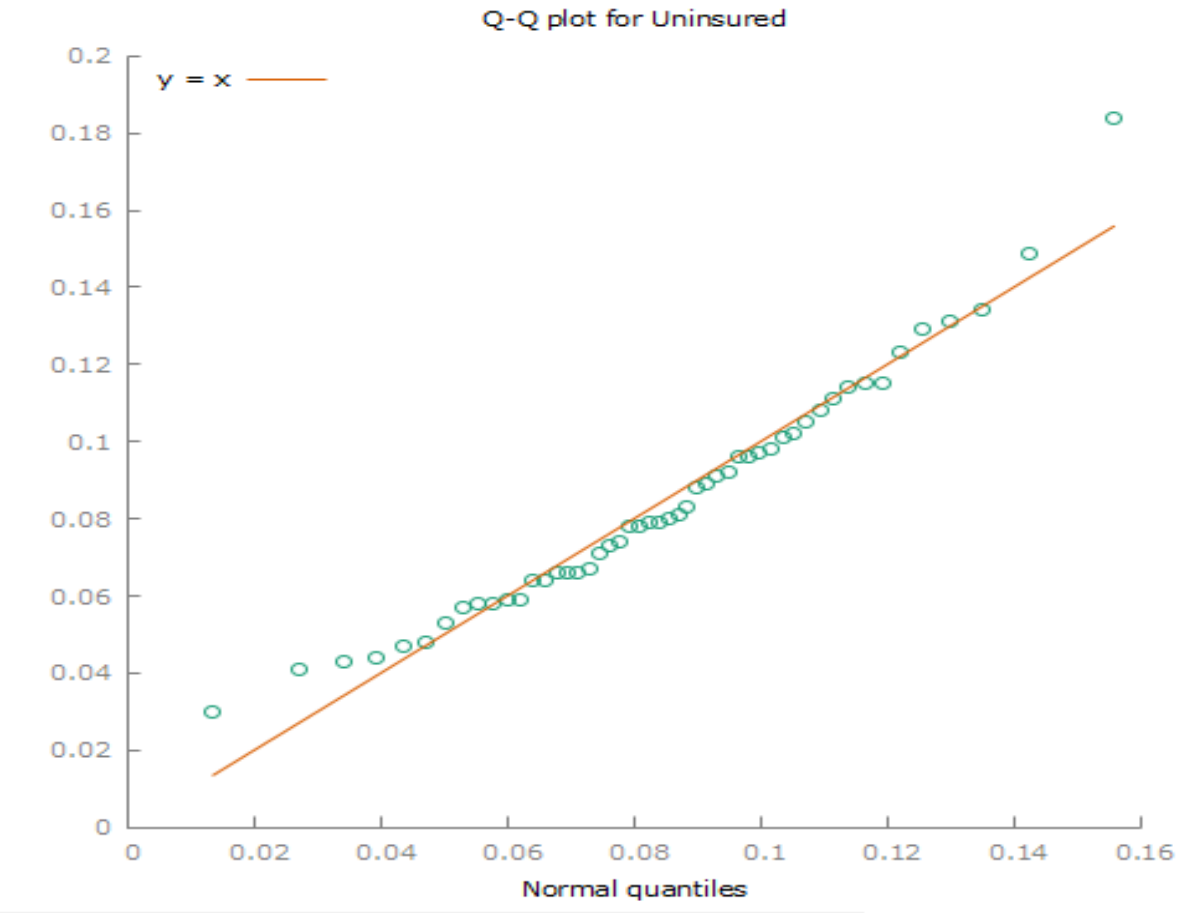
Mfrom50 QQ plot:



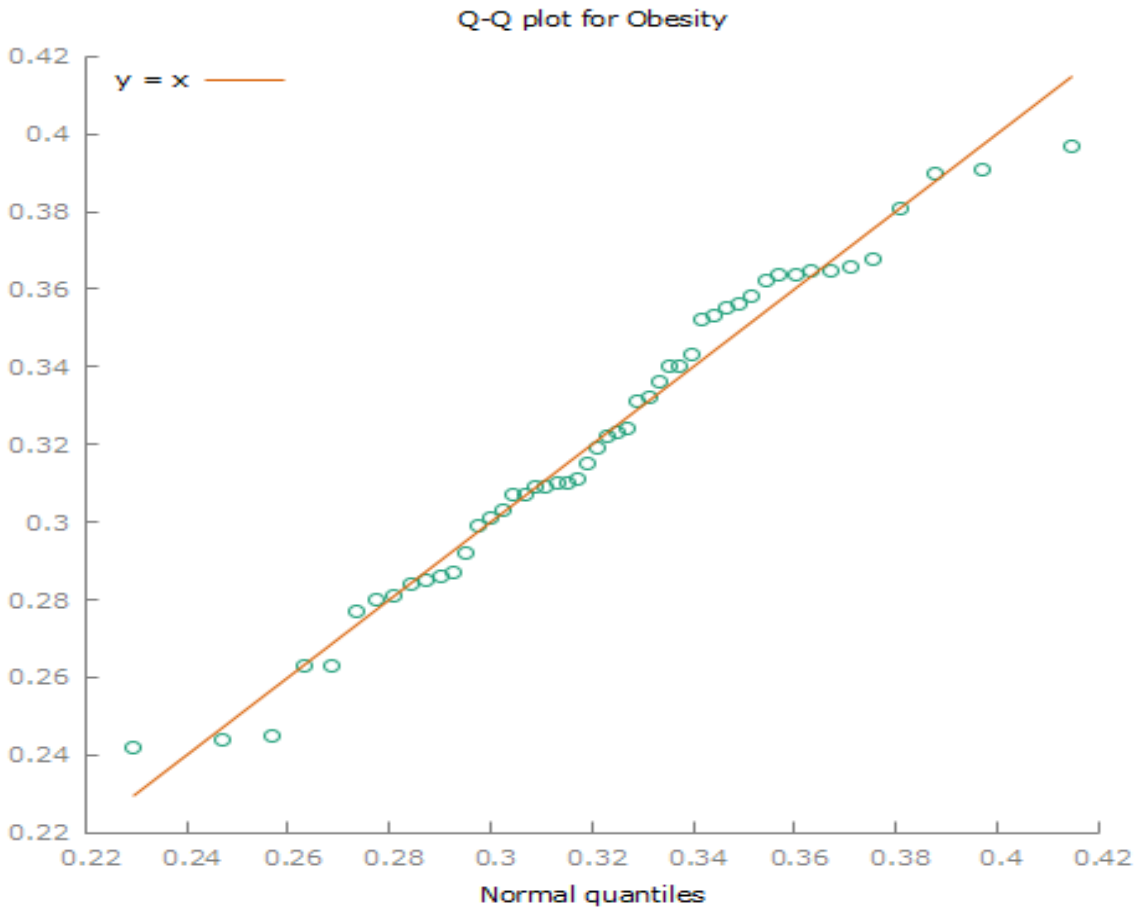
Smokers QQ plot:



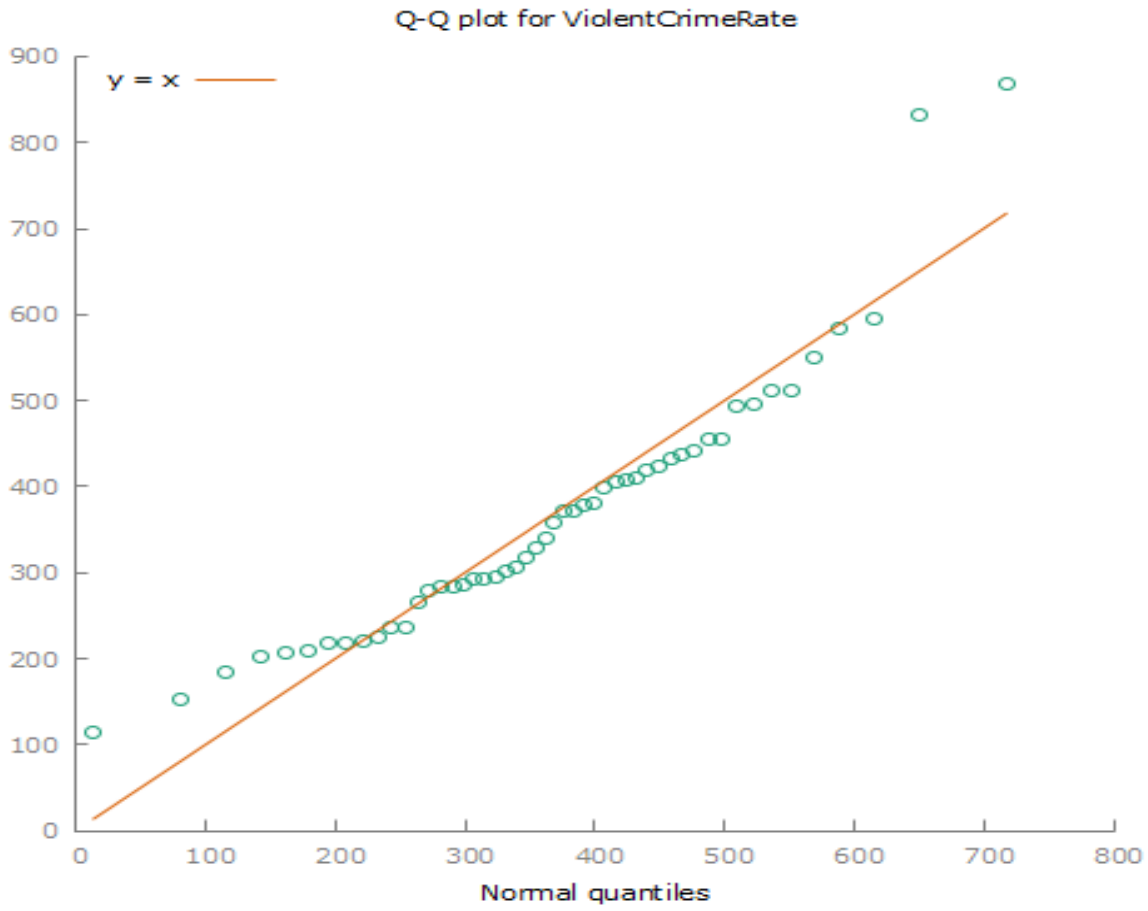
Uninsured QQ plot:



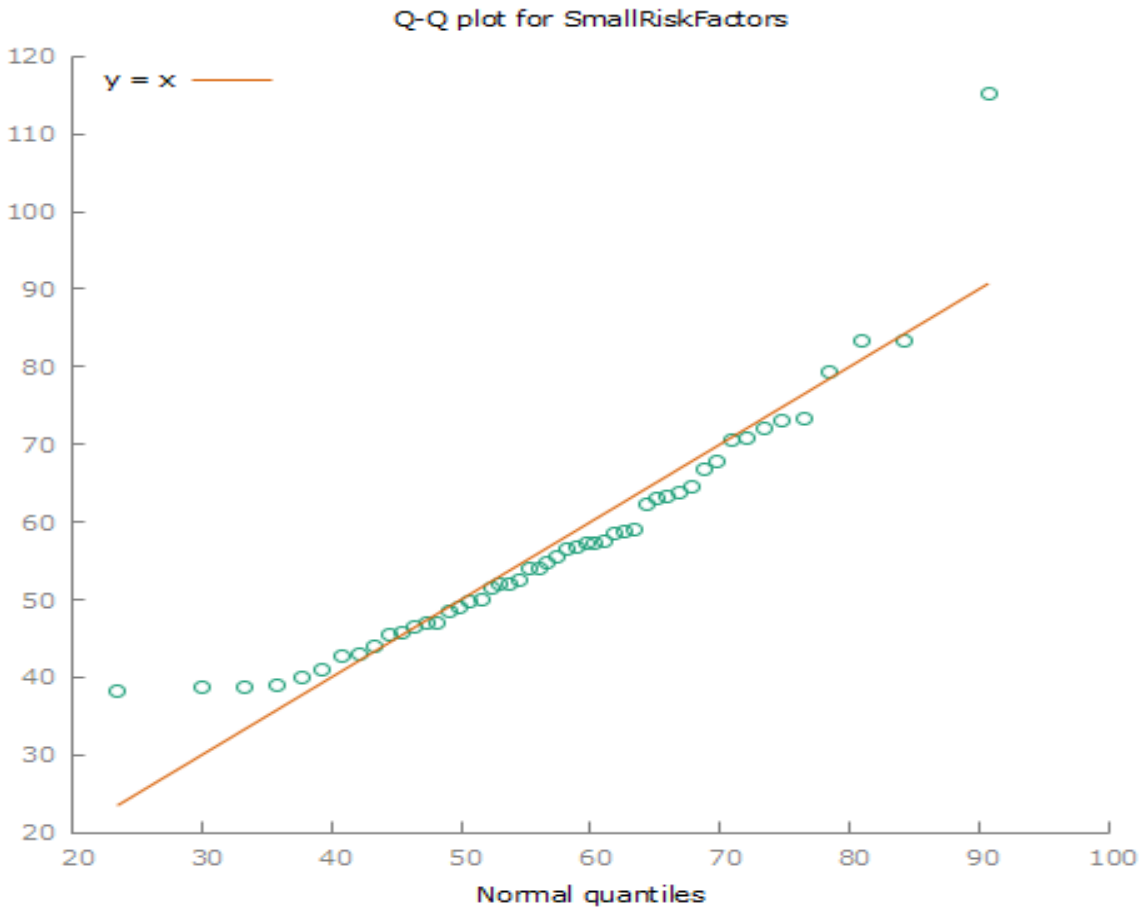
Obesity QQ plot:



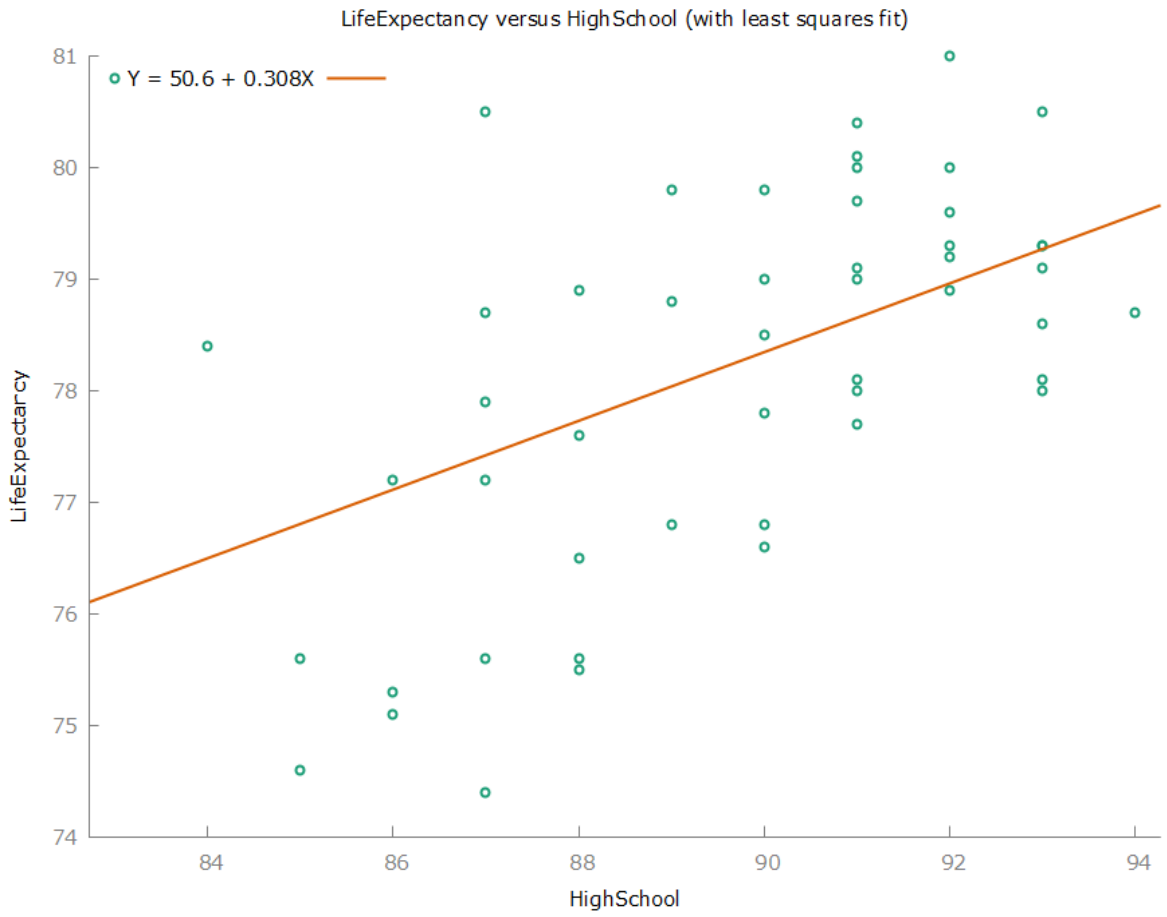
ViolentCrimeRate QQ plot:



SmallRiskFactors QQ plot:



Highschool QQ plot:



MHI1k QQ plot: