Exploring BMI Trends Amidst the COVID-19 Epoch: Predicting Changes and Examining Disparities Through Linear Regressions

Author: Faheem Khawar

In the last few decades, the world has seen an incredible increase in the production of food. Not only has the food supply exponentially increased, but it has also become relatively reliable and stable, along with the foods available being more diverse. However, a tradeoff had to be made between the quality of food and its quantity. Currently, American society has become extremely reliant on high-calorie nutrient-poor foods. Due to the COVID-19 pandemic, many of the issues involving this diet have been supposedly exasperated, leading to huge portions of the population into unfortunate decisions when it comes to their diet due to a multitude of different constraints. This begs the question, however, of how the average BMI of the population has changed between 2019 to 2023 and how to predict one's BMI given their characteristics and attributes.

In recent decades, societies have become dependent on processed foods generally considered unhealthy. Such foods usually are high in (saturated) fat and/or high in (added) sugar. Additionally, while offering high calories, they are poor in key vitamins and minerals such as potassium, fiber, and vitamin D. To add to this fact, these processed foods are often bought ready-to-eat from stores and are generally cheaper to buy than healthier foods. The cost needed to make nutritious meals consisting of fruits, vegetables, and high-quality carbohydrates (carbohydrates not stripped of their nutrients, fiber, and bran) has increased and poses a problem to people who don't have the time to prepare such meals when unhealthy alternatives are offered cheaper and are deemed as tastier and more filling. Due to being able to produce these foods on a massive scale and in a ready-to-eat form (such as chip bags and candy bars), many convenience stores have put them among their products for sale. Concurrently, items like fruits and vegetables can usually only be found in places like grocery stores, farmers' markets, and supermarkets, all of which are much fewer in numbers and less accessible compared to convenience stores.

Due to the advantages of buying foods considered unhealthy, in that it is generally cheaper, tastier, higher calories, and more accessible, such foods have become an integral part of many people's diets, to the point that people have become reliant on it. This is especially the case for those living in low-income and/or rural neighborhoods. Additionally, a situation referred to as a "food desert" can occur, where access to affordable fresh high-quality foods becomes challenging and difficult. A diet consisting mainly of these cheap processed foods, however, has shown disastrous consequences over a long period. Due to the nature of these high-calorie nutrient-poor foods, people have become deficient in vital vitamins and minerals while simultaneously being in a calorie surplus, resulting in weight gain. Additionally, this diet has also been shown to result in overconsumption of saturated fats, sugar, and sodium. Over-consuming these specific nutrients has been strongly associated with health problems such as obesity, hypertension, cardiovascular disease, and type II diabetes. Such health

problems have life-long implications and affect both the individual and society overall negatively.

This kind of diet has ultimately caused not just America, but many other first-world industrialized countries as well, to become fatter. The surplus of these foods has made populations increase in weight gain to the point where it is now becoming a health concern. It should also be noted that a concern like this on such a level is the first in known history, where instead of starvation and calorie deficiencies being the issue, it is now the overconsumption of calories and the issue of obesity.

A method experts have developed to calculate the overall health of a person is by using an individual's weight and height. The formula differs depending on which measurement system one uses, however, the formula for calculating body mass index, or BMI, under the imperial measurement system is shown below:

$$BMI = \frac{Weight\ (lb)*703}{Height^2\ (in^2)}$$

After observing the BMIs of the general population and statistically plotting said BMIs to see the distribution, the ranges of BMI are categorized under the following categories: BMI under 18.5 is classified as underweight, 18.5 - 24.9 is classified as normal/healthy weight, 25 - 29.9 is classified as overweight, and BMI above 30 is classified as obese. Due to the unhealthy diet, many Americans have adopted, the BMI of the general population has been believed to increase significantly.

With COVID-19 happening shortly after the start of 2020, many issues and inequalities across multiple domains (social, economic, political) have become more visible and in some cases, worsened. Additionally, the negative effects of the pandemic, like the mass shutdown of businesses, along with the policies and efforts to stabilize said negative effects have caused increases in inflation, unemployment, and cost of living. A major concern among this vast list of concerns was the (change in the) diet of the American people. Specifically, people would now likely become more dependent on places like convenience stores to obtain food for themselves, especially amidst the rising costs and shutdowns of many markets. This paired with the fact that many people were now forced to stay at home and eventually attempt to perform work remotely in a sedentary fashion led to a problematic issue where it seemed the general population was getting less physical activity alongside a poorer diet. Such a combination would lead to weight gain, putting people at higher risk of a multitude of health concerns, such as hypertension. A purpose of this paper, then, is to examine how the effects of the pandemic and the efforts made to combat them could have had on the overall BMI of the population, by state. In addition to this, it also plans to examine the relationship between BMI and income, race, education, and sex groups, along with unemployment.

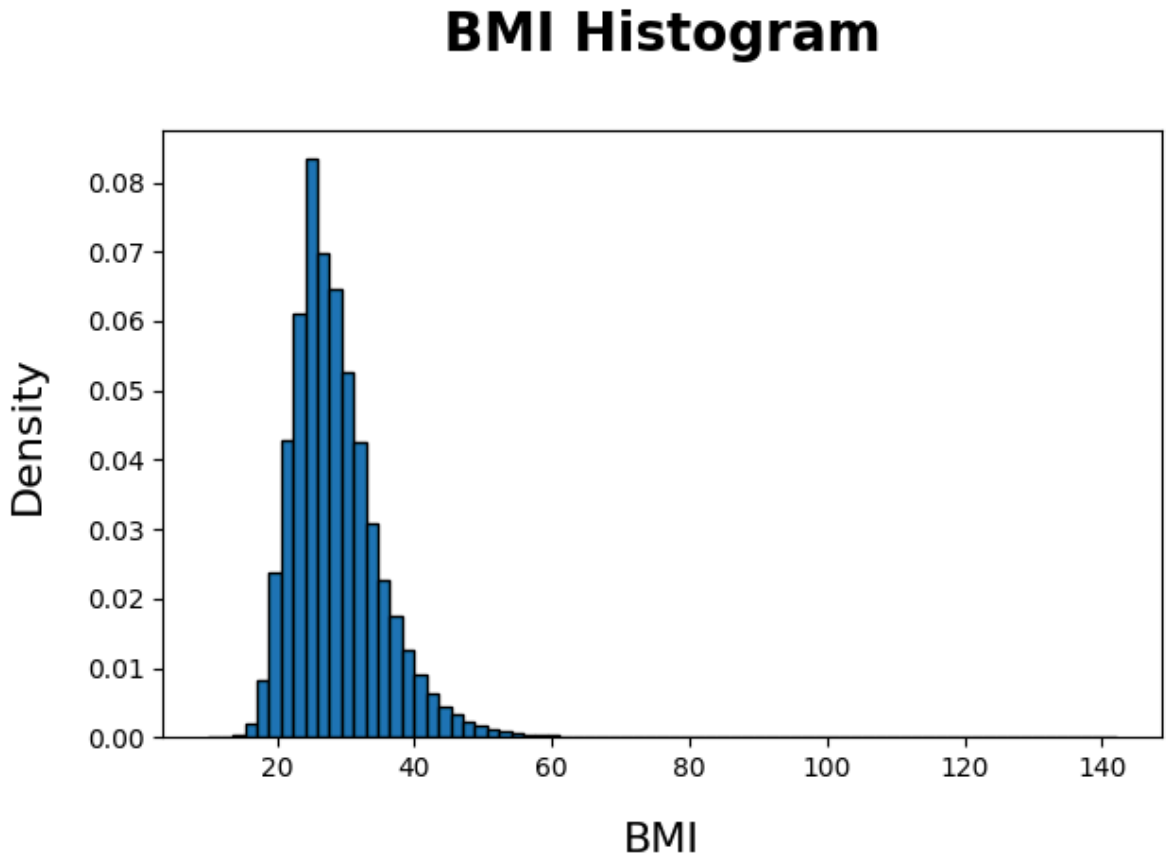
The importance of what this paper intends to do cannot be understated. In a society where it seems the overall population is already much heavier and unhealthier compared to the past, examining a potential worsening of this issue has drastic implications for every facet of society. To list a few, a heavier and unhealthier population means that the healthcare industry is now stretched to attempt to properly respond to the public. The overall population becomes less productive, leading to less economic growth and perhaps even economic decline. People with unhealthy lifestyles and considered overweight have also been associated with worse mental and

psychological health and even shorter lifespans. Being able to see what socioeconomic groups are dealing the most with (changes in) BMI, therefore, is vital to properly begin to respond to the issue. As society also continues to exit the COVID-19 pandemic, fully understanding the effects it has had on the population is imperative to not just fixing any negative effects, but to developing and growing as a society. Finally, seeing how unemployment is associated with (changes in) BMI allows society to know what to expect when times of economic trouble arise.

Data Gathering:

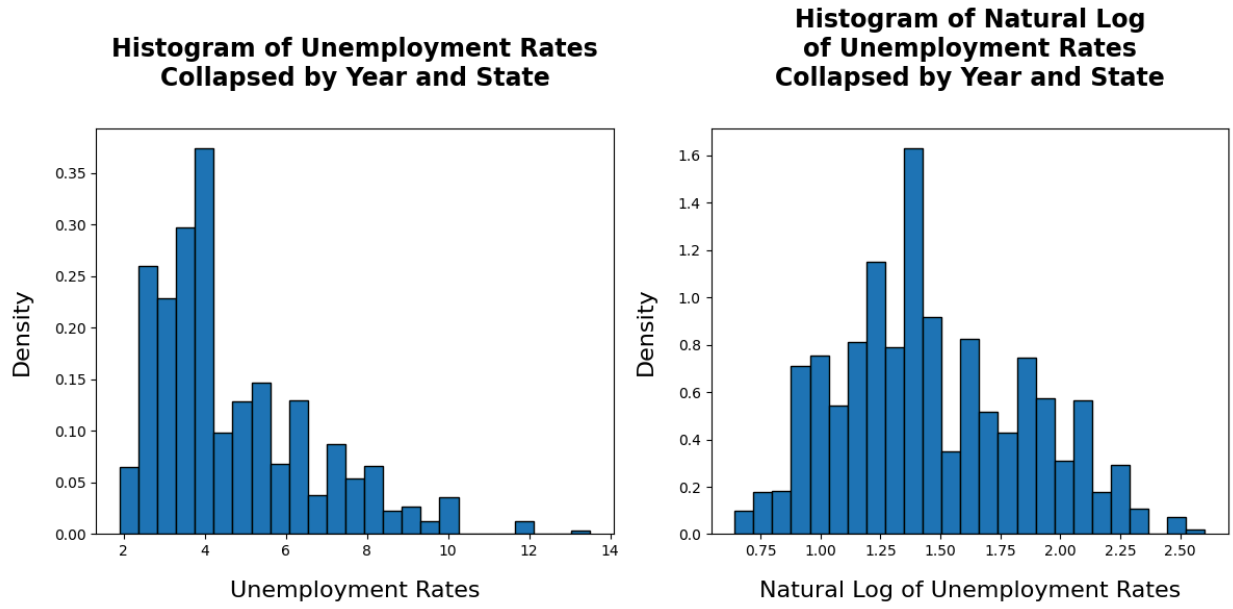
The data used in this study was mainly retrieved from the Behavioral Risk Factor Surveillance System, a subsection of the Center for Disease Control and Prevention. Each year was individually retrieved from their annual survey data (with the exception of 2023, which was grouped with 2022). The data for each year was then appended to a whole dataset, and each observation was on the individual level - each row represented an individual and their characteristics. The BMI was calculated by plugging in each individual's weight (in pounds) and height (in inches) into the BMI formula. Finally, data was retrieved from the U.S. Bureau of Labor Statistics, under Local Area Unemployment Statistics. The data for each year (within which the annual unemployment rates of each state) was gathered and merged with the original dataset. Additionally, the data removed any observations that gave no response to any of the variables of interest, along with removing the regions of Guam and Puerto Rico from the data (as the unemployment rates of these two regions were unavailable). Finally, any person with a calculated BMI of <10 or >150 was removed, as a BMI that high or low is nearly impossible (for BMIs >150, nearly all the heights of individuals were around 3 feet tall, it does not seem physically possible given the height and weight needed to achieve such a BMI).

Data Visualization:

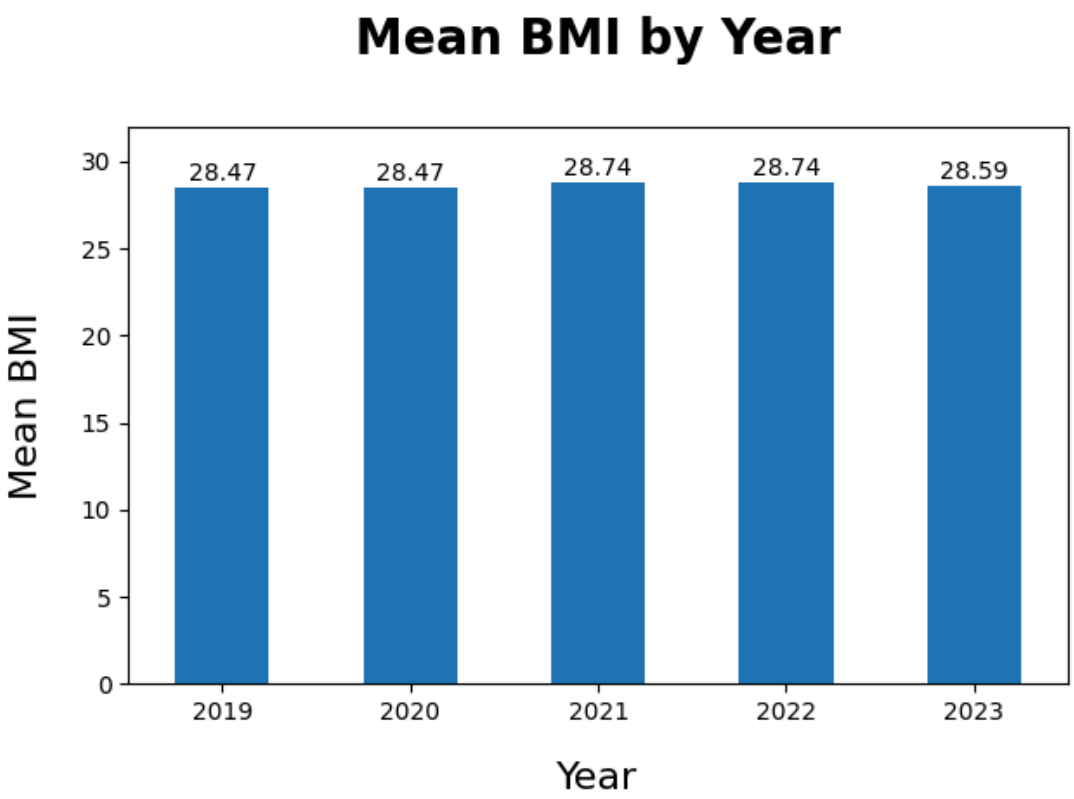


The histogram shows a relatively normal distribution of BMI values. While the distribution is slightly skewed right and does have outliers, it does not seem skewed enough to merit the use of taking the natural log of all the BMI values. To support this point, there is a fair possibility that some of the outliers, especially those towards the end of the scale, have incorrect BMIs, which is due to an error in reporting either weight, height, or both. This claim is supported by the fact that upon viewing some of the highest BMIs, many of the observations were extremely heavy while also being around 3 feet tall, something that seems fairly unlikely, especially given the number of observations that this trait. However, it was decided to keep such observations in the dataset for two reasons. First, it is possible that these observations are legitimate, and second, the effect keeping or removing said observations will have on the regression is minuscule. It is also impractical to sort through a dataset with more than one million observations to judge which observations seem invalid. Additionally, the goal of this paper is to be able to predict an actual BMI given a person's socioeconomic characteristics, the year, state, and unemployment level, and view changes across differences in variables (like year or state). Taking the natural log of BMI would make this goal difficult as the regression would now be talking in terms of percent change in BMI, which is not desired. Thus, the paper will assume a normal distribution of BMI and will proceed further with the regression under this

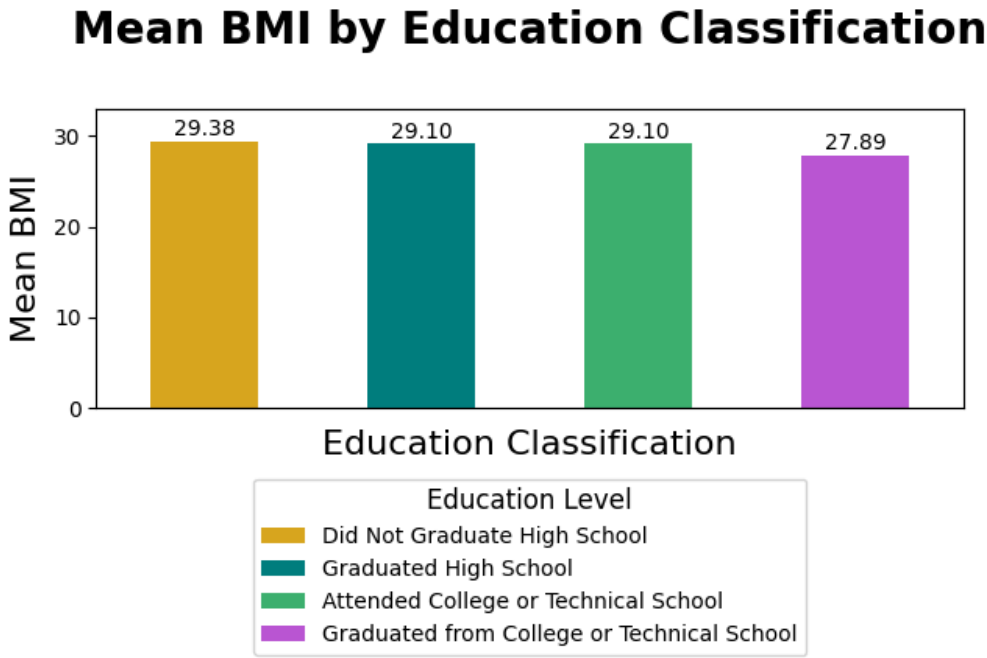
claim.



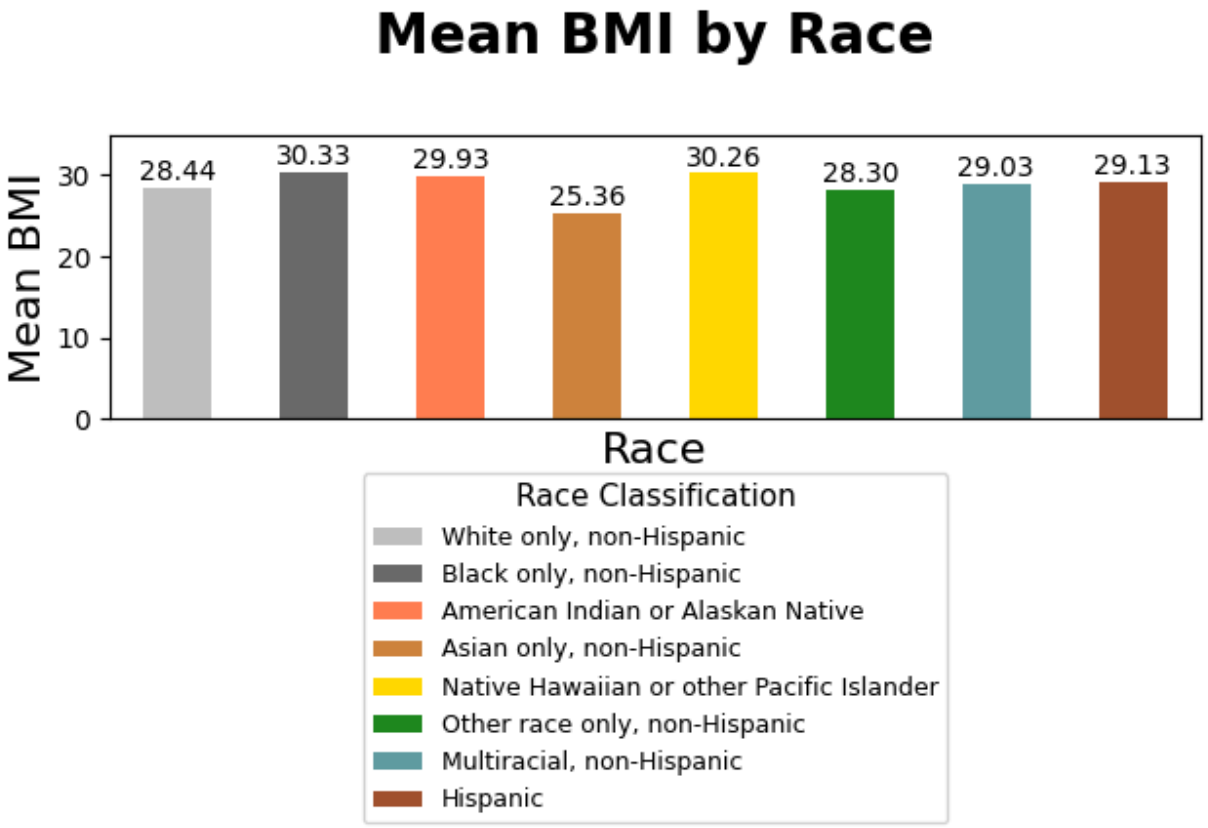
These two histograms are the unemployment rates of America when the data is collapsed by year and state (not on the individual level), to make the visualization of the unemployment rates easier to read. It was decided to take the natural log of the unemployment rates as by doing so, the distribution becomes much more normal compared to the distribution of unemployment rates, which is very right-skewed. Thus, going forward the regression will consider the rates in the natural log form.



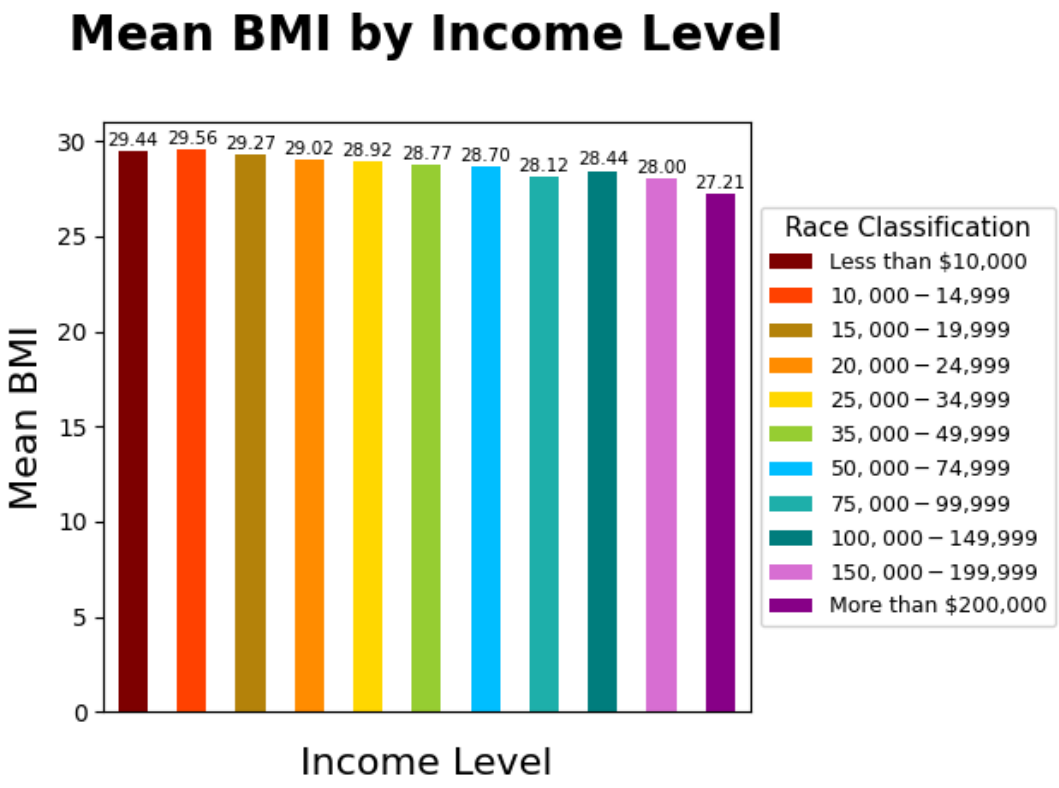
This bar chart shows the mean BMI for every year in the dataset. Taking into account standard deviations (which are similar in size between all the years), the mean BMI's are very similar.



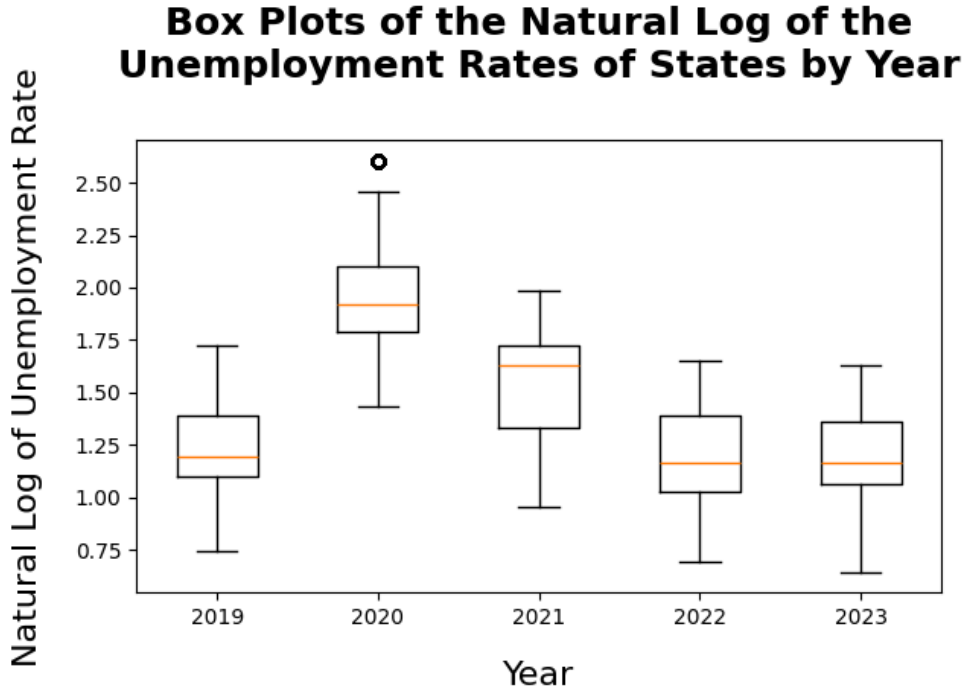
This bar chart shows the mean BMI of every education classification. It can be noted that as one's level of education increases, it appears that their BMI is expected to decrease as well.



The "mean BMI by Race" bar chart shows clear discrepancies between races. Races that are black, American Indian or Alaskan Native, and Native Hawaiian or other Pacific Island have the highest BMIs, while the races white and Asian have the two lowest BMIs, with Asians having an average of more than 3 BMI points lower than that of whites.



The contents of the "Mean BMI by Income Classification" bar chart show the average BMI by Income class. There appears to be a slight decrease in BMI as the income classes get higher (as one makes a higher income).



The image above shows the boxplots of the natural log on the state unemployment rates by year. It is clearly apparent that 2020, the year COVID-19 hit the world, has a much higher range than any of the other years. While the entirety of this huge increase in unemployment rates might not be completely due to COVID-19, it is likely that a fair portion of this increase is due to the pandemic (and any policies relating to it). Whatever factors that caused this increase also seemed to affect 2021 to a lesser extent, however, upon entering 2022 it appears that the unemployment rates are what they used to be (compared to 2019). The huge rise in unemployment rates and the huge differences between 2020 and the other years become important later when running regressions on the data.

```
Information on 'ln(unemployment rate)' variable:

count    1.229464e+06
mean     1.465354e+00
std      3.867319e-01
min      6.418539e-01
25%      1.163151e+00
50%      1.410987e+00
75%      1.757858e+00
max      2.602690e+00
Name: ln(unemployment rate), dtype: float64

Information on 'unemployment rate' variable:

count    1.229464e+06
mean     4.679377e+00
std      1.956176e+00
min      1.900000e+00
25%      3.200000e+00
50%      4.100000e+00
75%      5.800000e+00
max      1.350000e+01
Name: unemployment rate, dtype: float64
```

The summaries of the natural log of unemployment rates and unemployment rates are given (the rates are given without the natural log as it is more easily interpretable). It should be noted, however, as shown by the box plots above, that the mean of these rates by year differ significantly.

```
Information on 'BMI' variable:

count    1.229464e+06
mean     2.860523e+01
std      6.510136e+00
min      1.004075e+01
25%      2.420799e+01
50%      2.743572e+01
75%      3.179316e+01
max      1.420996e+02
Name: BMI, dtype: float64
```

The final piece of data visualization is the summary of the BMI, which includes the mean and standard deviation. The key takeaway from this summary is the mean, 28.61. Earlier, it was stated that the range of 25.0 - 29.9 is considered overweight, and anything equal to or above 30 is considered obese. From the dataset, it is revealed that the average BMI of an individual in America is within the overweight range. This should be extremely concerning to health agencies and the federal government, as an overweight population has many dire and adverse effects on society, as explained earlier. Additionally, not only is this mean BMI within the overweight range, but it is also a mere 1.39 points away from being considered obese, further emphasizing the problem.

Interpreting Changes in BMI:

Before going into the coefficient outputs of the regression, is it important to understand the value of the coefficients, which describe the changes in BMI. How big is a 0.3 increase in BMI? Is it big enough to matter, is it practically significant? While these answers, and the interpretation, are subjective and vary based on the individual, it is important to know how much a change in BMI actually means in terms of one's weight. To demonstrate this, an example will be made on an individual whose height and initial weight is the average of the height and weight of the dataset, which, rounded to the nearest whole number, is 184 pounds and 67 inches (5 feet 7 inches) resulting in a BMI of 28.8. Assuming one's height does not change and stays constant (a realistic assumption for an adult) an increase of 0.1 BMI points (leading to a BMI of 28.9) would mean that this individual would gain approximately 0.65 pounds. An increase of 1 whole BMI point would change this individual's weight by 6.4 pounds. It should again be noted that these increases in weight due to their respective BMI increases are not constant across individuals, a person who is 55 inches and 120 pounds will experience a different weight gain due to a 0.1 increase in BMI.

The individual used is just to give a rough idea of what the change in weight will be for the "average" person in our data. Again, how one interprets an increase of 0.65 pounds (from a 0.1 BMI increase) is subjective, however, given the discussion earlier on how the average person's BMI from our data is already classified in the "Overweight" category and being close to the "Obese" category too, an increase in weight should be worrisome. Additionally, while the change in weight associated with a small change in BMI may seem small, the magnitude of the changes can add up. These two points should be kept in mind while reading and interpreting the regression and its implications.

Regression Model:

$$BMI_{qtires} = \beta_0 + \beta_1 \ln ln\,(Unemployment\ Rate)_{st} + Sex_b + Year_t + Income_i + Education_e + State_s + \mu_{qtires}$$

BMI_{qtires} is the predicted BMI of an individual of state *s* in year *t* of sex *q* and race *r* given their income level *I* and education level *e*

$Unemployment\ Rate_{st}$ is the rate of unemployment in year *t* and state *s*

Sex_q is the sex fixed effect

$Income_t$ is the income fixed effect

$Race_r$ is the race fixed effect

$Education_e$ is the education fixed effect

$State_s$ is the state fixed effect

μ_{qtires} is the error term

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Main Linear Regression for this project:

OLS Regression Results

Dep. Variable:	BMI	R-squared:	0.028
Model:	OLS	Adj. R-squared:	0.028
Method:	Least Squares	F-statistic:	471.4
Date:	Wed, 14 Aug 2024	Prob (F-statistic):	0.00
Time:	16:57:51	Log-Likelihood:	-4.0301e+06
No. Observations:	1229464	AIC:	8.060e+06
Df Residuals:	1229387	BIC:	8.061e+06
Df Model:	76		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	29.3252	0.088	334.209	0.000	29.153	29.497
ln(unemployment rate)	0.1064	0.055	1.924	0.054	-0.002	0.215
state_Alaska	-0.6190	0.080	-7.785	0.000	-0.775	-0.463
state_Arizona	-0.9912	0.066	-15.052	0.000	-1.120	-0.862
state_Arkansas	-0.2481	0.073	-3.409	0.001	-0.391	-0.105
state_California	-1.3436	0.070	-19.207	0.000	-1.481	-1.207
state_Colorado	-1.6609	0.064	-26.072	0.000	-1.786	-1.536
state_Connecticut	-0.8086	0.068	-11.966	0.000	-0.941	-0.676
state_Delaware	-0.0963	0.081	-1.189	0.234	-0.255	0.062
state_District of Columbia	-2.0557	0.087	-23.559	0.000	-2.227	-1.885
state_Florida	-0.8360	0.063	-13.175	0.000	-0.960	-0.712
state_Georgia	-0.4308	0.066	-6.563	0.000	-0.559	-0.302
state_Hawaii	-1.4875	0.069	-21.613	0.000	-1.622	-1.353
state_Idaho	-0.5580	0.070	-7.990	0.000	-0.695	-0.421
state_Illinois	-0.5095	0.080	-6.359	0.000	-0.667	-0.352
state_Indiana	0.1723	0.064	2.672	0.008	0.046	0.299
state_Iowa	0.2408	0.063	3.801	0.000	0.117	0.365
state_Kansas	0.0789	0.060	1.305	0.192	-0.040	0.197
state_Kentucky	0.4234	0.074	5.686	0.000	0.277	0.569
state_Louisiana	-0.0670	0.075	-0.890	0.373	-0.215	0.081
state_Maine	-0.5337	0.063	-8.534	0.000	-0.656	-0.411
state_Maryland	-0.4524	0.059	-7.615	0.000	-0.569	-0.336
state_Massachusetts	-1.2561	0.068	-18.392	0.000	-1.390	-1.122
state_Michigan	-0.2277	0.068	-3.369	0.001	-0.360	-0.095
state_Minnesota	-0.3824	0.059	-6.536	0.000	-0.497	-0.268
state_Mississippi	0.1153	0.076	1.517	0.129	-0.034	0.264
state_Missouri	0.0762	0.064	1.190	0.234	-0.049	0.202
state_Montana	-0.8017	0.068	-11.842	0.000	-0.934	-0.669
state_Nebraska	0.1097	0.061	1.801	0.072	-0.010	0.229
state_Nevada	-0.9924	0.093	-10.708	0.000	-1.174	-0.811
state_New Hampshire	-0.6467	0.070	-9.256	0.000	-0.784	-0.510
state_New Jersey	-1.0048	0.075	-13.454	0.000	-1.151	-0.858
state_New Mexico	-1.1629	0.074	-15.737	0.000	-1.308	-1.018
state_New York	-0.8283	0.064	-13.027	0.000	-0.953	-0.704
state_North Carolina	-0.5195	0.075	-6.895	0.000	-0.667	-0.372
state_North Dakota	-0.0021	0.074	-0.028	0.977	-0.147	0.143
state_Ohio	0.3541	0.062	5.703	0.000	0.232	0.476
state_Oklahoma	0.1263	0.071	1.781	0.075	-0.013	0.265
state_Oregon	-0.7045	0.074	-9.540	0.000	-0.849	-0.560
state_Pennsylvania	-0.3906	0.074	-5.306	0.000	-0.535	-0.246
state_Rhode Island	-0.8236	0.073	-11.280	0.000	-0.967	-0.681
state_South Carolina	-0.4059	0.066	-6.118	0.000	-0.536	-0.276
state_South Dakota	-0.0894	0.070	-1.285	0.199	-0.226	0.047
state_Tennessee	0.0719	0.073	0.990	0.322	-0.070	0.214
state_Texas	-0.1695	0.064	-2.658	0.008	-0.294	-0.044
state_Utah	-0.7219	0.063	-11.417	0.000	-0.846	-0.598

state_Vermont	−1.0195	0.068	−14.978	0.000	−1.153	−0.886
state_Virginia	−0.3455	0.063	−5.480	0.000	−0.469	−0.222
state_Washington	−0.6942	0.062	−11.176	0.000	−0.816	−0.572
state_West Virginia	0.6004	0.073	8.275	0.000	0.458	0.743
state_Wisconsin	−0.0461	0.067	−0.684	0.494	−0.178	0.086
state_Wyoming	−0.8535	0.075	−11.327	0.000	−1.001	−0.706
year_2020	−0.0473	0.043	−1.101	0.271	−0.131	0.037
year_2021	0.3323	0.024	14.014	0.000	0.286	0.379
year_2022	0.4231	0.018	23.665	0.000	0.388	0.458
year_2023	0.3305	0.051	6.480	0.000	0.231	0.430
sex_2.0	−0.2125	0.012	−18.144	0.000	−0.235	−0.190
income_2.0	0.2201	0.044	4.976	0.000	0.133	0.307
income_3.0	−0.0647	0.041	−1.574	0.115	−0.145	0.016
income_4.0	−0.2253	0.039	−5.776	0.000	−0.302	−0.149
income_5.0	−0.3158	0.037	−8.555	0.000	−0.388	−0.243
income_6.0	−0.3264	0.036	−8.993	0.000	−0.397	−0.255
income_7.0	−0.2505	0.036	−6.971	0.000	−0.321	−0.180
income_8.0	−0.4871	0.035	−13.749	0.000	−0.556	−0.418
income_9.0	−0.3969	0.040	−9.809	0.000	−0.476	−0.318
income_10.0	−0.7166	0.047	−15.123	0.000	−0.809	−0.624
income_11.0	−1.3426	0.048	−28.242	0.000	−1.436	−1.249
race_2.0	1.7230	0.023	74.514	0.000	1.678	1.768
race_3.0	1.2620	0.046	27.613	0.000	1.172	1.352
race_4.0	−2.4323	0.040	−60.266	0.000	−2.511	−2.353
race_5.0	2.3036	0.101	22.811	0.000	2.106	2.502
race_6.0	−0.0120	0.078	−0.155	0.877	−0.164	0.140
race_7.0	0.6663	0.040	16.537	0.000	0.587	0.745
race_8.0	0.7528	0.024	31.525	0.000	0.706	0.800
education_2.0	−0.0210	0.029	−0.732	0.464	−0.077	0.035
education_3.0	0.1115	0.029	3.869	0.000	0.055	0.168
education_4.0	−0.7988	0.029	−27.450	0.000	−0.856	−0.742

Omnibus:	383149.453	Durbin-Watson:	1.988
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2061999.187
Skew:	1.405	Prob(JB):	0.00
Kurtosis:	8.688	Cond. No.	132.

The Stata output shown above is the regression used to predict the BMI of a given individual. The predictors of BMI for the regression were (the natural log of) the unemployment rate (of a given state in a given year), the individual's sex, race, income level, education level, the state of residence, and the year. All of the predictor variables, barring the unemployment rate, were fixed effects. It should be noted that while the sex variable was a fixed effect, the variable could have easily been implemented using a dummy variable and achieved the same results. Having the sex variable as a fixed effect was to clearly show which sex (male or female) corresponded to the coefficient in the output (in this case, being male was associated with a 0.215 increase in BMI). The base case of this regression is an individual from Alabama who is female, did not graduate from high school, receives an income less than $10,000 in the year 2019, is white, and faces an unemployment rate of zero. The classifications described are also the omitted classifications for their respective variable (for example, the omitted group of the state variable is Alabama). Additionally, all the coefficient outputs under each variable are relative to the omitted classification of said variable (for example, the -0.56 decrease in BMI in Idaho is relative to Alabama, the omitted state).

From the coefficient outputs of the states of America (including the District of Columbia), there is a pattern that the Southern states generally have a higher BMI than that of the Northern and Northeastern states. This aligns with the belief that the population of the southern states is, on average, unhealthier compared to the other regions. As for the coefficients of the education level, comparing the education levels to not graduating from high school shows little

difference in change in BMI and even a statistically insignificant coefficient for “Graduated High School.” However, the last education level, “Graduated from College or Technical School,” shows a huge BMI difference compared to not graduating from high school. The t-statistic to represent this difference is a staggering -27.45. The coefficient on sex showed that males have, on average, a higher BMI than women. This aligns with the fact that males are expected to have a higher BMI than females due to biological differences (like males have more muscle mass than females/females biologically have more body fat than males).

Upon looking at the coefficient outputs of the race variable, when the other races are being compared relative to the “White, non-Hispanic” race, there are clear discrepancies. The races of black, American Indian or Alaskan Native, and Native Hawaiian or other Pacific Island all have significantly higher BMIs than that of white people. This finding supports the extensive research showing inequalities between races in America, and that some races are, overall, unhealthier due to a multitude of factors (some of which will be described shortly). It is worth noting that while genetic and biological differences can (and do) cause individuals to gain more weight compared to others, the huge difference in BMI found here likely cannot be the sole reason for it.

The coefficient outputs of the income class, which are all relative to the income class earning less than $10,000, support the claim and existence of food deserts and food insecurity. Food insecurity “is defined as a household-level economic and social condition of limited or uncertain access to adequate food” (OASH Healthy People 2030). It also means that a household does not have enough stable access to food to support an active and healthy life(style). Food deserts, as explained earlier, are areas that have a lack of healthy and affordable foods, which are often found in areas with high levels of poverty, such as low-income neighborhoods or rural communities. From the coefficients, it is shown that those in the income class of $10,000 - $14,999 (and by extension and intuition, those earning less than $10,000) face a predicted increase in BMI. Potential reasons for this are that due to food insecurity and food deserts, individuals are forced to buy cheaper low-quality processed foods over nutritious ones, leading to a surplus in calories (leading to weight gain) while also being malnourished in key vitamins and minerals (like potassium). The income level of $15,000 - $19,999 could potentially be included in this discussion as, according to the regression, it not only has a mere expected decrease of -0.06 in BMI relative to those earning less than $10,000, but this value is not statistically significant. It should also be noted that low-income neighborhoods are disproportionately occupied by African Americans, along with other races. In the context of this regression, it could imply that one of the reasons the race “Black, non-Hispanic” faces a relatively high increase in BMI is due to the fact that a fair proportion of people in this group are in low-income communities facing food insecurity and living in food deserts (despite the best efforts of the fixed effects to separate the effects of BMI between race and income level). Finally, as the income level increases, the expected decrease in BMI seems to get bigger in value, implying that having more money allows one to spend on nutritious foods and meals and to maintain a healthy lifestyle, which intuitively makes sense.

The coefficient on the natural log of the unemployment rate, 0.106, is interpreted as follows: for every 1% increase in the unemployment rate, an individual’s BMI is expected to increase by 0.106%, or by 0.00106. Intuitively, one would believe that an increase in the unemployment rate would cause a higher increase in BMI, for reasons such as the amount of

stress would likely rise and financial constraints causing a poorer diet. The reason why this number is fairly small, perhaps smaller than anticipated, will be explained shortly.

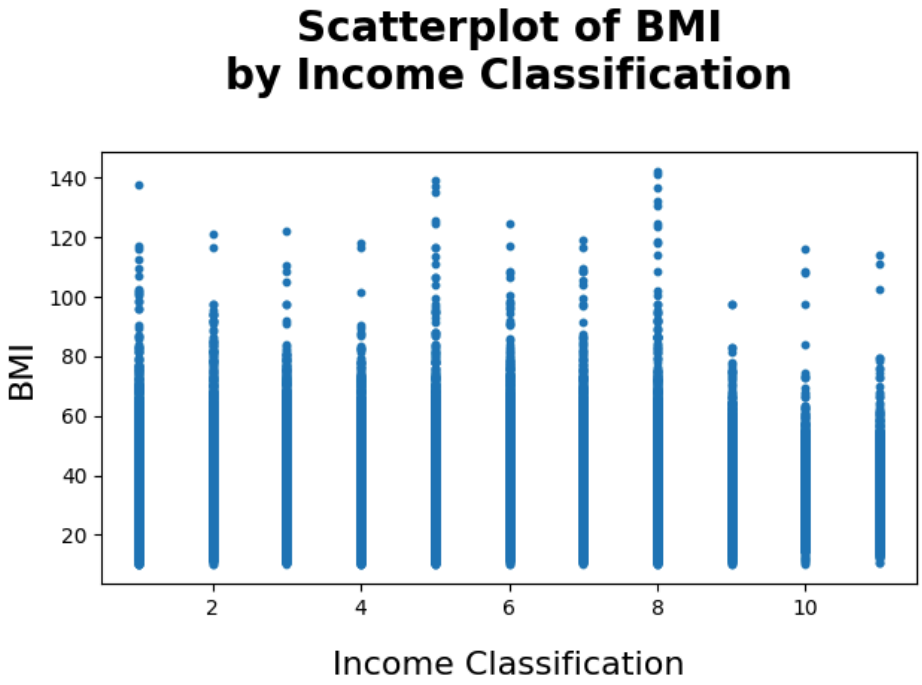
Finally, the last and arguably most important and telling variable of interest, year, had shown coefficients that are interesting, to say the least. Talking on the individual level, the other variables, like state and race, stay relatively the same between years (income class is included in this, both for simplicity's sake and due to the fact that it usually does not change between year to year). However, the year does not stay the same due to the passage of time (obviously), which means, in the context of this regression, one could predict how their BMI was expected to change simply from viewing the coefficient outputs associated with the years.

Looking first at the year 2020 (2019 is the omitted group, and all other years are relative to 2019), there is a predicted -0.047 decrease in BMI, additionally, recall that 2020 is also the year the COVID-19 pandemic hit. From this value, it could be interpreted that the pandemic potentially put a large portion of the population in situations that could cause weight loss (loss of financial means to purchase food, loss of access to stores (due to shutdowns), potential stress due to pandemic, etc.). However, this coefficient is not very large (near 0) and is not statistically significant. The more interesting coefficients lie for the years past 2020, both 2021 and 2023 have a BMI increase of 0.33, and 2022 has an increase of 0.42 (again, all years relative to 2019). These are fairly large increases in BMI across time and show that as society began to exit the pandemic and recover, the population seemed to undergo weight gain. But as society entered into 2023, the predicted BMI went down (going from a BMI increase of 0.42 in 2022 to an increase of 0.33 in 2023, both years relative to 2019). This decrease of approximately 0.1 points could show society beginning to stabilize into the “everyday life” that was pre-COVID-19. Putting this all together, a brief possible interpretation of the coefficients from the years variable is as follows: as the pandemic hit, there were no noticeable changes to the average person’s BMI. However, as the pandemic continued and entered into 2021, people’s lifestyle habits (and potential changes) started to catch up with them, causing an increase in weight. Entering 2022, this increase only worsened, but as society began to return to normal and entered 2023, people began to return to a healthier lifestyle causing a decrease in weight. How much of this weight change is being caused by the pandemic (and anything related to it) is unknown and would require a different study to go into it. It should also be noted that while the 2020 coefficient is relatively small, it's possible that an individual’s income level could have changed (due to being let go or a decrease in working hours) which, according to the regression, could have caused an increase in weight. The coefficient of 2020 being near 0 does not necessarily imply that the general population did not experience any weight change.

Also, to reemphasize again, while the interpretation and severity of these increases and decreases are subjective, seeing such an increase over the years in an already overweight population is concerning and could imply that the obesity crisis in America is at the very least still existent and not necessarily improving. One potential problem with this regression is that lifestyle choices, like diet, are not immediately recognized, and their effects may happen much later than the year in which the choice occurs – a time lag, in a sense. While it is entirely possible for such a scenario to be a (partial) factor for the coefficients of the years in this regression, properly exploring it is outside the scope of this paper.

Explaining R^2:

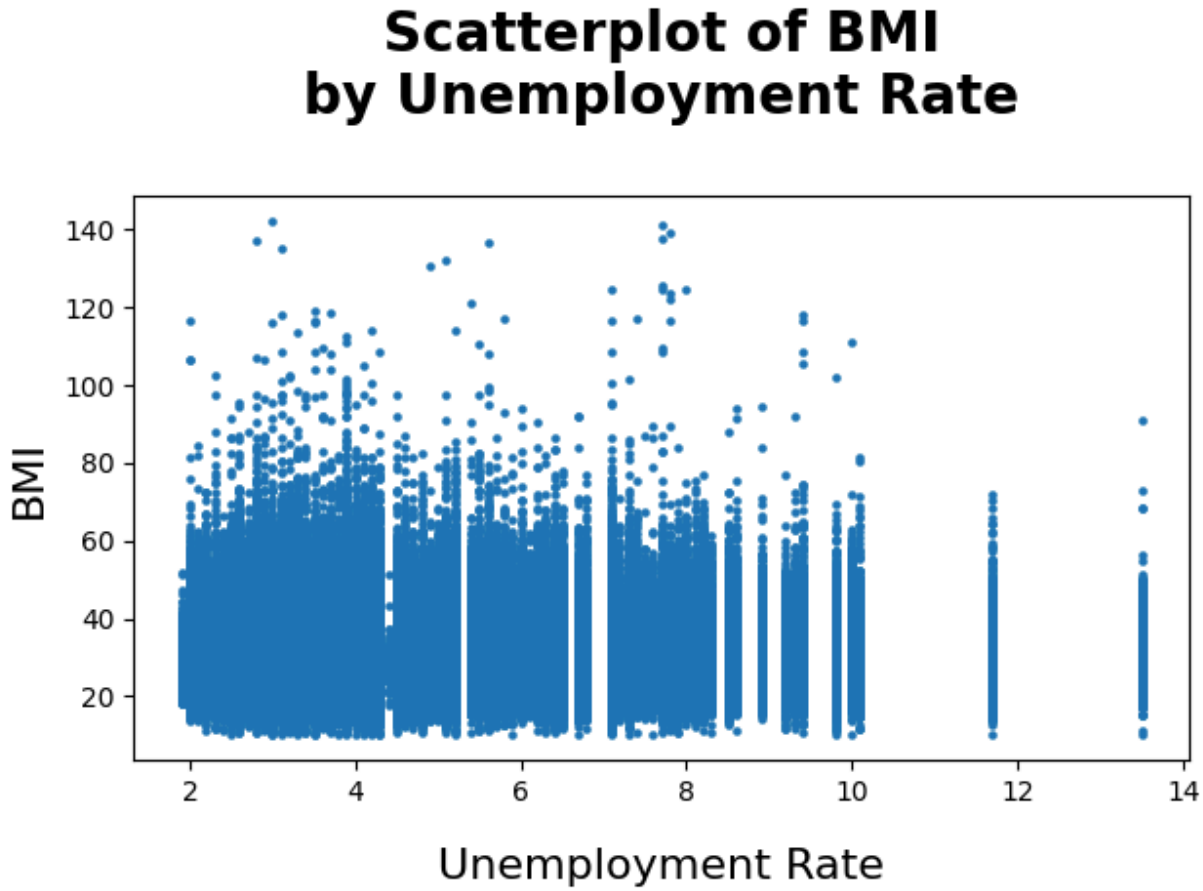
Upon examining the regression, one may quickly notice the shockingly low R^2 value, a measly 2.84%, or 2.83% for adjusted R^2. This percentage may imply that the regression model is an extremely poor predictor of BMI and thus not reliable, however, there is a legitimate explanation for such a low R^2. Barring BMI and the unemployment rate, all the other variables in the regression are discrete. The numerical values within each variable stand for some group, for example, the number 3 in the race variable stands for the "Asians, non-Hispanic" group. Due to this, when scatter plotting such variables with BMI, one would get something that looks like this:



Thus, due to the clustering and overlapping of data points, as well as the nature of discrete variables, a very poor line-of-best-fit is made, leading to a relatively poor R^2. Such a small percentage of the variation of BMI is explained by income levels (and nearly all the other variables) because of the discrete nature of said variables and due to how the data points are clustered.

What about the unemployment rates? Shouldn't a decent percentage of the variation in BMI be explained by these rates, especially since said rates shouldn't be discrete like the other variables? Not necessarily. Recall that this dataset is on the individual level (each row is an observation and its characteristics) not on a state-year level. This means that all people in a given state in a given year will have the same state-year unemployment rate, regardless of the variation between these individuals. For example, the scatterplot between BMI and unemployment rates is

as follows:



Much like before with the other variables, there is an over-clustering and overlapping of data points, leading to a poor line-of-best-fit and poor R^2. While this isn't ideal for making as accurate predictions as possible, having such a variable can still provide useful insight into predictions and still offer usefulness to the regression (something which will be argued shortly). The full implications and intricacies of including a variable of this nature in such a dataset, along with how it interacts with other variables, are beyond the scope of this paper, however, including such a variable is still valid and could offer useful information.

Alternative Regression Model:

Among the seven variables used to predict BMI, six of them describe a characteristic of an individual (like education, income, and state of residence). The only one that doesn't completely explain a characteristic is the unemployment rate. Additionally, due to the nature of fixed effects, one could argue that the fixed effects of the year variable (along with the state variable) should explain anything that the unemployment rates could explain as well, which is a valid point. The alternative regression that will be compared is the current regression minus the (natural log of) unemployment rates. To begin with, here is a regression of just BMI and the natural log of the unemployment rates:

Linear Regression of just ln(unemployment rate) and BMI:

OLS Regression Results

Dep. Variable:	BMI	R-squared:	0.000
Model:	OLS	Adj. R-squared:	0.000
Method:	Least Squares	F-statistic:	331.0
Date:	Wed, 14 Aug 2024	Prob (F-statistic):	5.99e-74
Time:	17:32:33	Log-Likelihood:	-4.0476e+06
No. Observations:	1229464	AIC:	8.095e+06
Df Residuals:	1229462	BIC:	8.095e+06
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	29.0099	0.023	1261.010	0.000	28.965	29.055
ln(unemployment rate)	-0.2762	0.015	-18.193	0.000	-0.306	-0.246

Omnibus:	388529.031	Durbin-Watson:	1.967
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2030524.973
Skew:	1.437	Prob(JB):	0.00
Kurtosis:	8.601	Cond. No.	8.41

Recall that from the original regression, the coefficient of the rates was approximately 0.106. This means that when including other variables like state and education level, the coefficient of the unemployment rates went down, making the regression of just BMI and the rates biased upwards. The coefficient for the unemployment rates being much lower in the original regression could mean that the fixed effects of other variables, namely year and state, help explain a lot of the effects these rates would have on BMI. Now, here is the original regression minus the rates (only sex and year of the regression will be shown, as there is very little difference between the other variables):

Alternate Linear Regression where unemployment rates are entirely excluded:

OLS Regression Results

Dep. Variable:	BMI	R-squared:	0.028
Model:	OLS	Adj. R-squared:	0.028
Method:	Least Squares	F-statistic:	477.6
Date:	Wed, 14 Aug 2024	Prob (F-statistic):	0.00
Time:	16:57:41	Log-Likelihood:	-4.0301e+06
No. Observations:	1229464	AIC:	8.060e+06
Df Residuals:	1229388	BIC:	8.061e+06
Df Model:	75		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	29.4393	0.065	455.485	0.000	29.313	29.566
year_2020	0.0288	0.017	1.720	0.085	-0.004	0.062
year_2021	0.3630	0.018	20.705	0.000	0.329	0.397
year_2022	0.4164	0.018	23.746	0.000	0.382	0.451
year_2023	0.3209	0.051	6.322	0.000	0.221	0.420
sex_2.0	-0.2125	0.012	-18.143	0.000	-0.235	-0.190

Omnibus:	383151.763	Durbin-Watson:	1.988
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2062021.989
Skew:	1.405	Prob(JB):	0.00
Kurtosis:	8.688	Cond. No.	89.1

The coefficient on sex is nearly the same, however, the main difference lies in the year variable. While the coefficients for the years 2022 and 2023 stay nearly the same and 2021 changed by 0.03, the coefficient for the year 2020 is completely different. When including the unemployment rates the coefficient is -0.051, but when excluding the rates it is now 0.029 (rounded), a near 0.08 difference. This change could be explained by the fact that while year-fixed effects can account for the factors that are different across units but constant over time, they cannot account for factors that are both different across units and across time. Recall from the data visualization about the unemployment rates how the year 2020 was significantly higher than the others. The shutdown caused by COVID-19 caused the unemployment rates of all of the states to increase, but COVID-19 did not hit all states to the same degree and magnitude. Some states had higher changes in unemployment rates compared to others, and the year-fixed effects could not account fully for this. These significant changes can be seen in 2020 and partially in 2021, as seen from the regressions and implied by the boxplots shown earlier. Thus, including the (natural log of) unemployment rates helped to account for these inconsistent changes. The coefficient for the year 2020 is biased upwards when not taking the unemployment rate into consideration in the regression.

A solid counterargument can be made that not including the unemployment rates in the regression would not make a significant difference, and perhaps be more accurate. Notice how in both versions of the regressions, the coefficients on the year 2020, the only group across all groups in the regressions to change a fair amount, are not statistically significant and their 95% confidence intervals include 0. Not only that, but the coefficient for 2020 in the alternative regression has a higher t-statistic compared to the original regression and actually is statistically significant at a significance level of 10%. While insight and predictions can still be drawn from coefficients that are not statistically significant (that is to say that whether a coefficient is statistically significant is not the be-all-end-all), the legitimacy of including the unemployment rates due to the change in coefficients for 2020 can most certainly be challenged.

Short Comings:

While this paper used BMI to measure body fat and to roughly determine overall health, it should be worth mentioning the criticisms and shortcomings of using this index. To begin with, the BMI struggles to differentiate body fat from muscle mass. Due to this, people with a lot of muscle mass (like bodybuilders or athletes) are often reported and classified as overweight or obese under the BMI system despite being extremely healthy. While there are certainly some individuals in the dataset who fit this description, most people would not suffer from this shortcoming of the BMI system. So for the most part, this shortcoming shouldn't affect the results of the regression to a noticeable degree.

A much more problematic shortcoming of the BMI is that it is criticized for fundamentally being a poor indicator of general health, and perhaps even misleading. This is due to the fact that along with not taking into consideration muscle mass, it fails to account for things like bone density biological differences (between sexes and races), and overall body composition. There is also a critique that due to how height is taken into account equationally by squaring it, it causes BMI to overestimate "fatness" in tall people and "thinness" in short people.

These criticisms of the index are certainly valid, so much so that it is important to make this clear: the BMI is a flawed metric. However, that does not mean it cannot give meaningful insight into the overall health of the general population. While BMI cannot, and should not, be

the sole indicator of someone's health, it certainly can be a very rough indicator of health and can be a supplemental method used to determine health. Additionally, it is much easier and cost-effective to use BMI to help determine health, as it only requires weight and height, two measurements included in nearly every medical report and measurements every person usually knows about themselves off the top of their head.

Other methods of measuring body fat that do not have some of the shortcomings of the BMI are methods such as using the body adiposity index, measurements of visceral fat, and waist circumference. It should be noted that these methods are either more costly, more difficult to preform, or both compared to getting an individual's BMI. The argument can be made, however, that using BMI as the indicator of health in this regression is misleading and inaccurate, which is a legitimate standing given the BMI's fundamental shortcomings. Perhaps running a regression to predict weight, taking into account height, would have been more appropriate. Regardless, the paper takes the stance that the BMI is a valid index to use to *roughly* judge the overall health of an individual and the population along with observing health discrepancies among groups as well as changes in health based upon the BMI system.

Conclusion:

From the results of the regression, it can be inferred that as the years progressed, the overall BMI (and thus weight) of the American population increased until 2023, when the average weight seemed to slightly decrease. It was also found that the average BMI of an individual in America is approximately 28.8, which is considered overweight and close to being considered obese. As the unemployment rate increases, so does one's predicted BMI. Males are shown to have higher BMIs than females, which can at the very least be partially explained by biological differences between both sexes. Southern states are shown to be unhealthier due to their higher BMI averages than other regions (like the Northern or Northeastern states). Additionally, it seems that the higher one has in terms of education, the lower their BMI is. Some races seem disproportionately higher in their BMIs than other races, most noticeably blacks, American Indians or Alaskan Native, and Native Hawaiian or Pacific Island. It should also be noted that these races, especially blacks, are more likely to live in worse-off and low-income neighborhoods than other races, like Asians. These discrepancies, alongside the findings of the regression for the income variable support the claim that food deserts and food insecurity could be causing those with low incomes (earning less than $14,999, potentially those earning less than $19,999) to have higher BMIs and be heavier than those with higher incomes.

The regression and data showed existing inequalities and disparities between different groups, the reason for which is complex and due to a multitude of factors. It also supports the notion that economic decline causes weight gain. Additionally, while the BMI seemed to decrease in 2023, keeping this decrease long-term is imperative to producing a healthier population and tackling the multiple problems and effects that come with an overweight population. The average BMI is still concerningly high and, along with the evident disparities of weight between different groups, need to be taken seriously to ensure that society can function at an efficient level. Means of tackling these problems vary, but some include community outreach and education, a decrease in the price of healthy produce and foods, and social programs to help those who need it.