North Carolina Journal of
Mathematics and Statistics
Volume 10, Pages 20–39 (Accepted May 6, 2024, published May 10, 2024)
ISSN 2380-7539

Mathematical Approaches of Modeling Obesity Trends

DANIELLE DASILVA AND KAREN YOKLEY

ABSTRACT. The prevalence of obesity has drastically increased over the past several decades and has caused strain within the healthcare system, as obesity puts individuals at an increased risk for a variety of diseases and conditions. This project develops multiple mathematical models for obesity trends in the United States. We first used linear regression to model how the overall trends of obesity have changed over time. Linear regressions enabled us to gain insight into the relationship between obesity and societal factors such as poverty and food insecurity and enabled us to gain insight into the relationships seen in the data. Further, the rise in obesity levels has been theorized to mimic the spread of an infectious diseases. Since infectious diseases are often studied using SIR-models, we next developed an SIR model to study and analyze their effectiveness in modeling obesity. This enabled us to gain an understanding of the population level dynamics however might be overly complex. Finally, we used agent-based modeling strategies to create a probabilistic model of obesity trends. The use of agent-based models is supported by the theory that one's social community may also impact the likelihood of becoming obese. The agent-based model was relatively simple but modeled the population level dynamics well. Developing these and similar models could enable the investigation of various intervention strategies to reduce obesity levels within the United States.

1. Introduction

Obesity, defined by the CDC (2022b) as having a BMI greater than 30.0, puts people at greater risk for conditions such as heart disease, stroke, type 2 diabetes, and certain types of cancers including pancreatic, liver, and kidney cancer. These conditions are among the leading causes of preventable, premature death in the United States (CDC, 2022a). The United States obesity prevalence has risen to 42.4% in 2017-2018 from 30.5% in 1999-2000 (Hales et al., 2020). This increase is putting a strain on the overall healthcare system as the estimated annual medical cost of obesity in the United States was nearly $173 billion in 2019 dollars. Medical costs for adults who had obesity were $1,861 higher than medical costs for people with healthy weight (Ward et al., 2021) Since obesity increases one's risk of getting severely ill from COVID-19 (Kompaniyets et al., 2021), the medical cost of obesity has likely increased in the past several years.

Mathematical models that describe the spread of infection in a population are well-established (Smith et al., 2004). Although obesity is not contagious, how obesity spreads through a population might mimic the trend of an infectious disease. Various researchers have looked at how to model obesity with epidemiological models. Santonja et al. (2010) analyzed the incidence of excess weight in adults in Valencia, Spain and analyzed how strategies such as healthy advertising campaigns could be an effective way of controlling the increase of adult obesity. Delavani et al. (2021) created a differential equation model to investigate how obesity spreads among the human

Received by the editors February 27, 2024.
2020 *Mathematics Subject Classification.* 92-10; 92B05.
Key words and phrases. SIR; Obesity; Mathematical Modeling; Population Dynamics.

population and the impact of media campaigns but did not compare the model to data. Thomas et al. (2014) aimed to model changes in BMI and the conditions under which obesity prevalence will plateau. Their model was based on data from 1988 to 1998 in the United States and the United Kingdom and considered both social and non-social influences on weight gain. However, they do not consider a wider variety of influences, such as income or education, and they do not compare different regions of the United States. Little research has focused on creating a mathematical model describing how societal factors influence one's likelihood to become obese, especially when looking specifically at adults in the United States. By better understanding the societal factors that are correlated with higher levels of obesity, healthcare professionals and policymakers can work to counteract these factors and reverse the increasing prevalence of obesity.

One societal factor that is often connected to obesity is poverty. Researchers theorize that an increase to access to healthy food for the nation's poorest would decrease obesity levels in the country. However, the research regarding the correlation between poverty and obesity have conflicting conclusions. Some researchers (such as Levine (2011)) suggest that poverty and obesity are correlated because individuals who live in impoverished regions have reduced access to fresh food due to phenomena such as food deserts. They observed that food insecurity along with greater sedentary behavior is correlated with higher rates of obesity. This could be due to the low cost of energy-dense foods, the decline in fruit and vegetable consumption or various psychological and behavioral changes, such as a preoccupation with food, stress, depression, and physical limitations in adults (Dinour et al., 2007). Ogden et al. (2010) have differing conclusions to their research. They conclude that among men, obesity prevalence is generally similar at all income levels, with a tendency to be slightly higher at higher income levels. Among women, higher income women are less likely to be obese than lower income women, but the majority of obese women are not low income. Low income was defined as having an income below 130% of the poverty level, and high income was defined as having an income at or above 350% of the poverty level (Ogden et al., 2010). Published research also suggests relationships between socioeconomic status and obesity and between race/ethnicity and obesity (Fryar et al., 2012). Comparing trends between different demographic factors and obesity levels across the United States over time will enable us to gain insight into the influence these demographic factors might be having on obesity trends.

Social influences are also theorized to be connected to obesity levels. Specifically, Christakis and Fowler (2007) show how obesity spreads through a social network over time. Christakis's study shows that having a friend who is obese increases a person's chance of becoming obese more than having a spouse who is obese. One's family environment may also impact one's likelihood to become obese in other ways, Hernandez et al. (2015) showed the impact of infant feeding methods on obesity levels. Additionally, Smith et al. (2020) found that obesity, weight, and dietary behaviors are influenced by one's social ties. Theorized mechanisms by which this occurs include social support, social norms, social comparison, or behavior modeling.

The current research project involved the creation and analysis of various models of obesity trends. These models incorporated the contribution of different factors, such as poverty, to the rising obesity trends and can be used to estimate the effectiveness and outcomes of intervention strategies. First, overall trends and factors were investigated through linear regression models. Linear regressions between obesity levels and societal factors such as poverty and food insecurity were conducted over time and by state. Next, compartmental differential equation models were analyzed to assess their effectiveness in modeling obesity. This project adapted an SIR model to describe the obesity trends throughout the United States population utilizing a system of three equations that is subsequently simplified to a system of two equations. Finally, agent-based models

were investigated to see if population trends within a defined group of individuals can be captured. The use of agent-based models is supported by the theory that one's social community may also impact the likelihood of becoming obese. Ideally, agent-based models could be adapted to reflect these trends. The goal of creating and analyzing these models is to accurately analyze the factors contributing to the rise in the prevalence of obesity in the United States, predict future trends, and develop strategies to minimize the prevalence of obesity.

2. Methods

According to the CDC (2022b), body mass index, or BMI, is a person's weight in kilograms divided by the square of height in meters. If a person has a BMI less than 18.5 they are considered underweight, between 18.5 and 25.0 they are considered healthy weight, between 25.0 and 30.0 they are considered overweight, and if a person's BMI is 30.0 or higher they are considered obese.

BMI can be used as a screening tool, but does not diagnose health. A trained health care provider should perform appropriate assessments to evaluate an individual's health status and risks. BMI is moderately correlated with more direct measures of body fat obtained from skinfold thickness measurements, bioelectrical impedance, underwater weighing, dual energy x-ray absorptiometry, and other methods. However, BMI appears to be strongly correlated with various adverse health outcomes consistent with more direct measures of body fat (CDC, 2022b). Despite obesity not being a diagnosis of health, for the duration of this paper, obesity will be compared to a healthy-weight category as a typical naming convention. This healthy-weight category is assumed to also include underweight individuals. The healthy weight percentages were calculated based upon the assumption that the population adds up to 100%, so the healthy weight percentage was assumed to be 100% minus the reported obesity percentage and the reported overweight percentage. Additionally, unless an overweight category is specified, overweight individuals have also been included in the healthy-weight category because the levels of overweight individuals have not significantly changed over the past ten years (CDC, 2023a).

The data used in this project were gathered through the Behavioral Risk Factor Surveillance System (BRFSS) (CDC, 2014). The BRFSS is a national system of health-related telephone surveys that collect state data about United States residents regarding their health-related risk behaviors, chronic health conditions and use of preventive services. The BRFSS completes more than 400,000 adult interviews each year. BMI was calculated from self-reported weight and height. Respondents weighing less than 50 pounds or more than 650 pounds, with height below 3 feet or above 8 feet and with a BMI less than 12 or greater than 100 were excluded from the obesity data used in this study. Pregnant respondents were also excluded from the data. Obesity data collection strategies have changed over time, and hence, for consistency, only data collected for 2011 or later was used.

Various software was used throughout the course of this project. The data were downloaded from the CDC website and then processed and cleaned into a usable format in Microsoft Excel and SQL. To perform the correlation analysis, the fit function within and differential equation analysis, MATLAB, version R2022b, was used (`https://mathworks.com`). For the multivariate linear regressions, Python through Google Colaboratory (`https://colab.research.google.com`) was used and the LinearRegression function was used within the linear_model class of the sklearn module (Pedregosa et al., 2011). To initially develop and visualize the differential equation model, Wolfram Mathematica 12 was used (`https://wolfram.com`). The numerical ODE solver, ode45, which is is based on an explicit Runge-Kutta (4,5) formula was used in MATLAB. The optimization toolbox within MATLAB was also utilized for estimating

parameter values within the differential equation model, specifically fmincon was used. FminCon finds the minimum of a constrained nonlinear multi-variable function utilizing an interior-point algorithm and allows the restriction of parameter values to be within a specified range. In order to develop the agent based model, NetLogo 6.3.0 was utilized (`https://ccl.northwestern.edu/netlogo/`).

In order to investigate the relationship between poverty and obesity, national and state level poverty data were retrieved and reformatted from the CDC's Chronic Disease Indicator website (CDC, 2024) for each of the years between 2011 and 2019. Obesity data were also gathered from the CDC's obesity data (CDC, 2023a) website for each state for each year between 2011 and 2019. Data for the poverty levels and the obesity levels for each of the 50 states in 2019 were gathered. Data for other societal factors were cleaned and prepared for analysis. One of these societal factors was the percentage of people meeting aerobic physical activity guidelines for substantial health benefits among adults aged 18 or older. The CDC further define this measure as the percentage of adults who reported at least 150 minutes per week of moderate-intensity physical activity, or at least 75 minutes per week of vigorous-intensity physical activity, or a combination of moderate-intensity and vigorous-intensity physical activity (multiplied by two) totaling at least 150 minutes per week (CDC, 2024). Data were also gathered from the USDA regarding food insecurity for each state in 2019 (USDA, 2023). Food insecurity was defined as households that were uncertain of having or unable to acquire enough food to meet the needs to all their members at times during the year because they had insufficient money or other resources for food. Next, the prevalence of sufficient sleep among adults aged 18 years or older was retrieved from the CDC (CDC, 2024). Sufficient sleep was defined as getting 8 hours or more sleep for those aged 18 to 21 and getting 7 hours or more sleep for those above 22 years old on average during a 24-hour period.

For each income bracket, obesity data were gathered. The CDC separates income brackets into less than \$15,000, between \$15,000 and \$24,999, between \$25,000 and \$34,999, between \$50,000 and \$74,999, and finally as \$75,000 and above. These income levels were self-reported and participants were asked for their annual household income from all sources (CDC, 2023a). Since the data were divided into income level instead of above and below the poverty line, categories of having an income less than \$15,000 and having an income of more than \$15,000 were used to define those in poverty and those not in poverty. However, since this is annual household income and the poverty line varies based upon household size, this designation is not completely accurate. Those with income not reported were excluded from the analysis. The reasons for unreported income were not specified. Among people with an income level less than \$15,000, 32.3% of people were obese in 2011 and 38.7% of people were obese in 2021. Among people with an income of more than \$15,000, 27.5% of people were obese in 2011, and 35.7% of people were obese in 2021. In the sample, the number of people with an income of less than \$15,000 decreased from 49,598 to 19,516 from 2011 to 2021. This raw data supports Levine (2011)'s conclusion that poverty is correlated to higher obesity levels.

3. Models

3.1. Correlation Analysis

Poverty has decreased over the past ten years in the United States. Obesity has increased over the same time period. The percentage of people in poverty in the United States was plotted in MATLAB for each of the years from 2011 to 2019, and then the percentage of people in the United States classified as obese was also plotted on the same graph over the same time interval.

These trends are illustrated in Figure 3.1. The inverse relationship that is seen in the data implies that either the decrease in poverty has a delayed impact on obesity, or the correlation between poverty and obesity is not as straightforward on a national level as researchers such as Levine (2011) suggest. The lack of correlation is potentially due to the fact that less than 20% of obese adults are at or below the poverty line. Thus, relatively small decreases in poverty would not lead to any significant changes in overall obesity levels because the vast majority of obese individuals are not in poverty (CDC, 2023a).

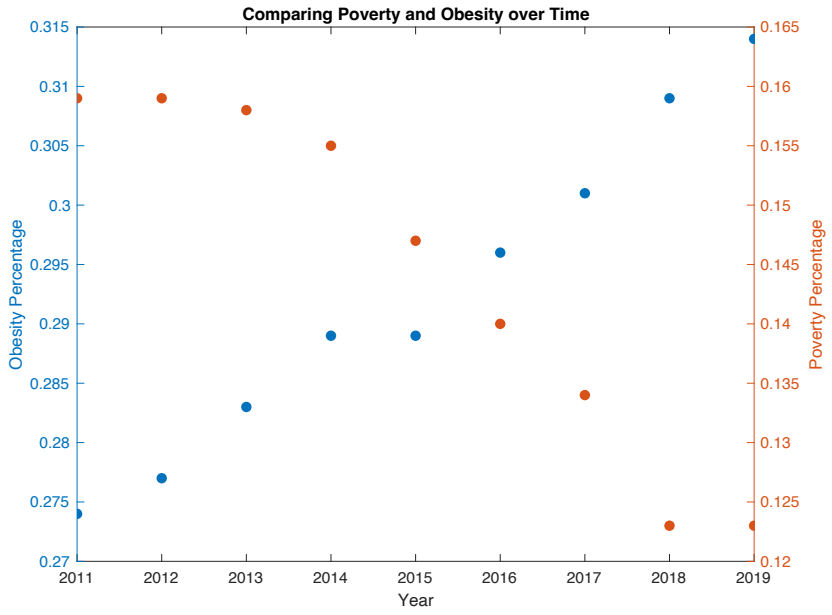


FIGURE 3.1. Comparison of poverty (CDC, 2024) and obesity (CDC, 2023a) trends from 2011 to 2019

For each of the states in 2019, the data were plotted with obesity levels on the x-axis and poverty levels on the y-axis. A linear regression between these factors was conducted and the strength and direction of the correlation was analyzed. In Figure 3.2, a weak correlation between poverty and obesity is observed when obesity levels and poverty levels in 2019 are compared across states. The R^2 value of this relationship is 0.283, implying that 28.3% percent of variance in obesity levels can be explained by the variance in poverty levels. This supports the claim that poverty is not the main or only driving factor of rising obesity levels across the United States.

The influence of other factors, such as exercise levels, food insecurity, and sufficient sleep on obesity trends was also explored on a state-by-state basis. A linear regression was then conducted and analyzed for each of these societal factors. These factors have R^2 values of 0.434, 0.318, and 0.2276 respectively. Thus, these factors do not have a very strong correlation with obesity levels on a population scale either.

A multivariate linear regression was then conducted in Python to further explore the relationship between various societal factors and obesity levels. In the multivariate regression, the explanatory factors that were analyzed included the percentage of people meeting exercise standards, experiencing food insecurity, meeting sleep guidelines, in poverty, and with sufficient health insurance. This multivariate linear regression specifically analyzed these factors in 2019 compared with the

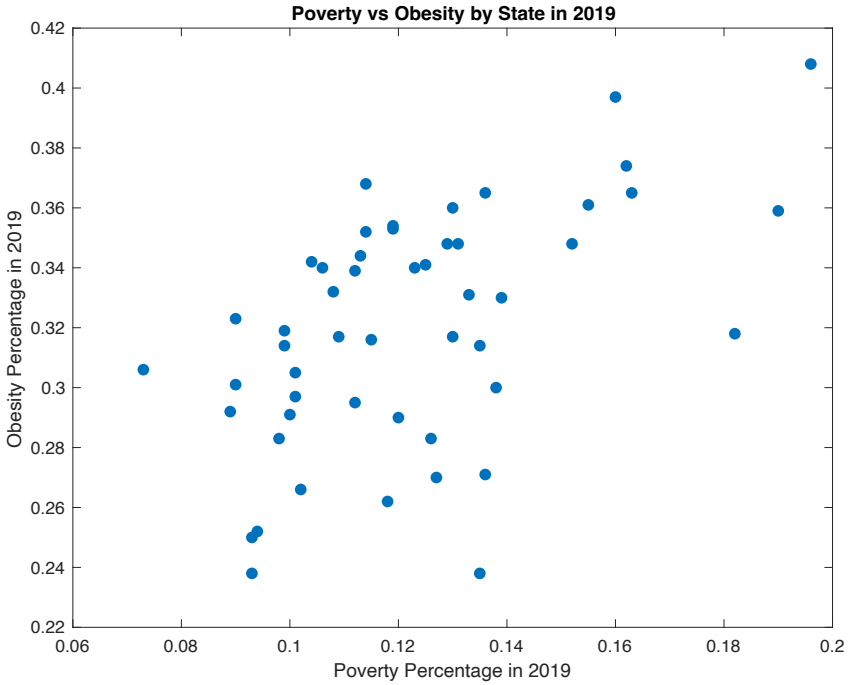


FIGURE 3.2. Comparison of poverty (CDC, 2024) and obesity (CDC, 2023a) levels in 2019 for each state

percentage of people with obesity. An R^2 value of 0.33 was discovered for this relationship meaning that 33% of the variation in obesity can be explained by the variation in these five factors.

For each of the CDC defined income brackets, a line was plotted in Matlab showing the obesity percentages for individuals in that income bracket over the years from 2011 to 2019. The trends for each income bracket were compared and analyzed. On a national level, the percentage of people who are obese and in poverty was compared to the percentage of people who are obese and not in poverty. Figure 3.3 shows that people in lower income brackets have higher levels of obesity (CDC, 2023a). However, this data also tells us that the average yearly increase in obesity is 0.64% for those in poverty (or with an income of less than $15,000) and the average yearly increase is 0.82% for those with an income of greater than $15,000. Thus, higher income individuals are becoming obese at a slightly higher rate. Additionally, those with an income not reported are greatly impacting the yearly increase of obesity levels, as the average yearly increase in obesity levels of the entire population is 0.56% which is lower than either of the two categories independently.

3.2. SIR Model

3.2.1. Methods

Previous researchers have attempted to mathematically model obesity utilizing an SIR-type model to capture obesity trends (Santonja et al., 2010; Delavani et al., 2021; Thomas et al., 2014). Thus, the next model developed in the current investigation was a three compartment differential equation model which incorporates societal factors. The impact of factors such as decreasing poverty

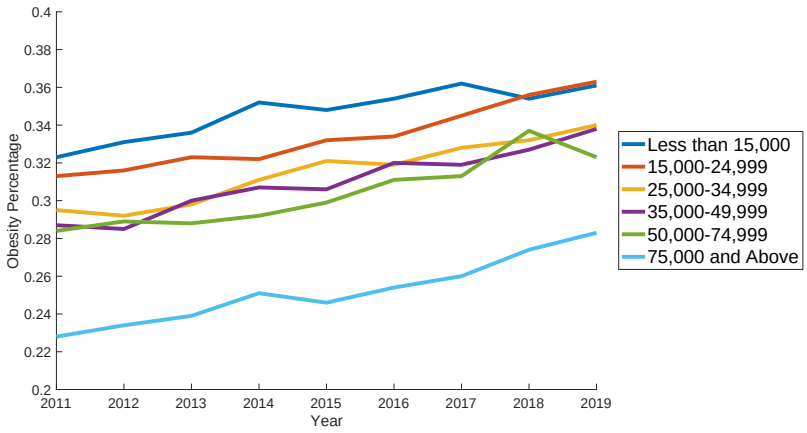


FIGURE 3.3. Obesity levels over time by income bracket

on obesity trends could potentially be analyzed. The three compartments are healthy weight, overweight, and obese. As in previous research, both transition and interaction terms were incorporated into the model. The model was defined by the following system of differential equations:

$$s'(t) = -\beta_1 * s(t) * o_v(t) - \beta_2 * s(t) * o_b(t) + r_1 * o_v(t) \tag{3.1}$$

$$o_v'(t) = \beta_1 * s(t) * o_v(t) + \beta_2 * s(t) * o_b(t) - k * o_v(t) + r_2 * o_b(t) - r_1 * o_v(t) \tag{3.2}$$

$$o_b'(t) = k * o_v(t) - r_2 * o_b(t) \tag{3.3}$$

where s is the healthy weight category, o_v is the overweight category, and o_b is the obese category and the parameters are defined in Table 3.1. This model is a simplification of the models by Delavani et al. (2021) and Thomas et al. (2014) which had systems of four and six equations respectively. The system is also a simplification of the model by Santonja et al. (2010) as it does not account for the movement of people in and out of the system and assumes one recovery parameter from obesity to overweight and the same recovery parameter from overweight to healthy weight. Previous researchers (Christakis and Fowler, 2007) have theorized that social interaction influences someone's likelihood to gain weight due to common shared behaviors and norms. People are assumed to transition in and out of obesity based on a transition term. This transition could result from two potential interactions that could cause someone to transition into the overweight class: (1) an interaction between a healthy weight person and an overweight person which is governed by a transition rate of β_1, and (2) an interaction between a healthy weight person and an obese person which is governed by a transition rate of β_2. Based upon simulations, it was theorized that β_2 would be approximately four times greater than β_1 in order to minimize our error functional. Two recovery rates, r_1 and r_2 were used in the creation of the model. However, when the model was optimized to the data, the resulting values of r_1 and r_2 were equal. Parameter optimization and sensitivity analyses were then conducted. During parameter optimization, β_1, β_2, and k were constrained to be optimized between 0 and 1 and r_1 and r_2 were constrained to be optimized between 0.00001 and 0.1.

Preliminary analysis suggested that the three compartment differential equation model was an over-complication of the data and the information available about obesity dynamics. The differential equation model was then simplified to have only two categories. The system of equations

TABLE 3.1. Table of States and Parameters

State	Description
s	healthy weight (percentage of population)
o_v	overweight (percentage of population)
o_b	obesity (percentage of a population)
o	combined overweight and obesity
Parameter	**Description**
β_1	transition rate due to interaction between s and o_v population
β_2	transition rate due to interaction between s and o_b populations
β	transition rate due to interaction between s and o populations
k	transition rate from overweight to obese
r_1	recovery rate from o_v to s
r_2	recovery rate from o_b to o_v
r	recovery rate from o to s
t	time (years)
N	total population

governing the model were defined as follows:

$$s'(t) = -\beta * s(t) * o(t) + r * o(t) \tag{3.4}$$

$$o'(t) = \beta * s(t) * o(t) - r * o(t) \tag{3.5}$$

where s is the healthy weight category, and o is encompassing of an overweight and obese category. Additionally, β is the rate of transition between healthy weight and overweight due to interaction, and r is a recovery rate, or the rate at which people transition from overweight to healthy weight. The interaction term is theorized to be a result of shared genetics with early life relationships, learned behaviors, and shared social norms.

In order to gain a proper understanding of the whole population, it was assumed that

$$N = s + o = 1 \tag{3.6}$$

where N is the total population as a percentage, s is the proportion of healthy weight individuals and o is the proportion of overweight and obese individuals. The o category is defined by adding the o_v and o_b categories in the previous system of three differential equations. This assumption enabled the utilization of overweight and obesity data from the CDC to compare to the model, as the data was in percentages that add up to 100%. It was assumed that the healthy weight population is equal to $1-(o_v+o_b)$. Published literature guided the ranges for the parameters. Previous research enabled us estimate r, or the rate at which people lose weight and transition from obesity to healthy weight, to be between $\frac{1}{124}$ and $\frac{1}{210}$ because researchers estimated that for people classified as obese, the probability of attaining a normal weight was 1 in 210 for men and 1 in 124 for women (Fildes et al., 2015).

The parameters in equations 3.4 and 3.5 were optimized for this model based upon the limited data and information available. An error functional was utilized which calculated the sum of the squared distance between the model prediction and the CDC data for each year. Values for both β

and r were optimized. First, we restricted r between 0.0048 and 0.0081, and then we optimized for β assuming that r is constant at 0.005.

Parameter values were then estimated separately for 49 states and the District of Columbia. New Jersey was excluded due to the fact that there was data missing for 2019 for New Jersey. For each state, β values were optimized for both variable r values and constant r values. The optimized β values for each state were then plotted against the previously mentioned societal factors in 2019, such as percentage of people in food insecurity, meeting exercise standards, in poverty, and meeting sleep guidelines. Both linear regressions and a multivariate linear regression were conducted to understand the correlation between β values and the societal factors on a state-wide level with the goal of being able to create an equation to estimate β based upon knowledge of societal factors.

3.2.2. Results

For the three compartment model, the data from the CDC were used to optimize the model parameters. On a national level, Figure 3.4 implies that the data aligns relatively well to the optimized model. The model adequately takes into account the trends of the data and how various categories are changing over time. The overweight population remains relatively constant while the obese population increases and the healthy weight population decreases. With the further optimization of parameters, the model could become better aligned to the data. Since the model only captures a small time frame for a slow progressing trend, the typical shape of an SIR-type model is not captured within the time frame shown. However, the parameter values were difficult to optimize based upon the data alone and the model was not sensitive to certain parameter values. This lack of sensitivity can be seen in Figure 3.5. It is seen that when varying β_1 and β_2 by up to 10 times the original values there is little impact to the obesity curve of the graph. However, for parameters that the model is sensitive to, such as k, varying those parameters by up to 10 times results in almost double the obesity values over the 10 years modeled.

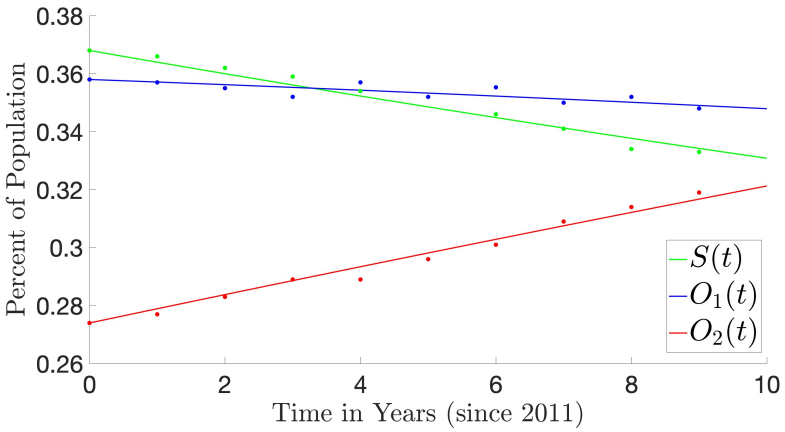
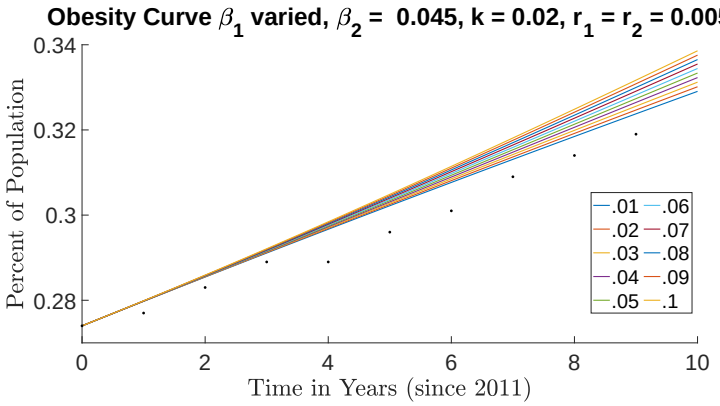
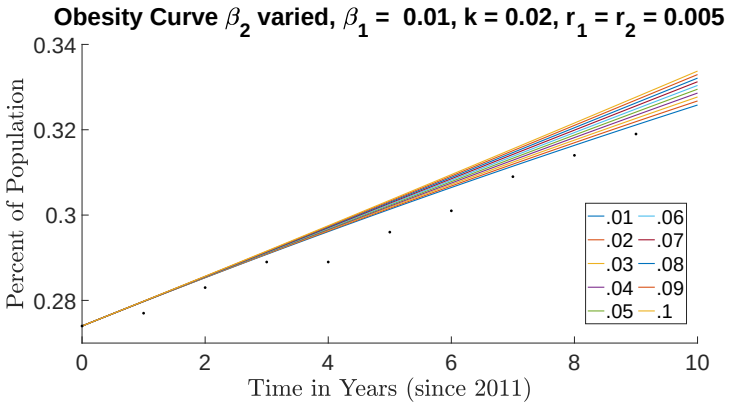


FIGURE 3.4. Solution curves for equations (3.1) - (3.3) compared to United States data (CDC, 2023a) with a $\beta_1 = 0.01$, $\beta_2 = 0.045$, $k = 0.02$, and $r = 0.005$.

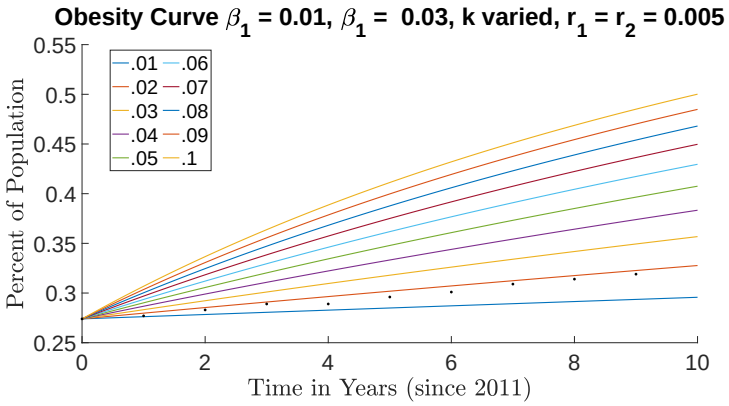
When the model was simplified from a three compartment to a two compartment model, the model was still able to fit the data (CDC, 2023a) extremely well between 2011 and 2019 as seen in Figure 3.6. The following optimal parameters were found with the specified error value when considering national data:



(A) β_1 sensitivity



(B) β_2 sensitivity



(C) k sensitivity

FIGURE 3.5. Obesity curve with parameters being varied compared to United States data (CDC, 2023a).

- $\beta = 0.0305$
- $r = 0.0049$
- error: 8.9623×10^{-5}

When r was restricted to be equal 0.005, the following optimal parameter was found with the specified error value:

- $\beta = 0.0307$
- error: 8.9814×10^{-5}

Thus, on the national level, keeping r constant can produce almost as low of error as letting r vary. These error values can be compared to the minimum error achieved on a national level for the system of three differential equations which was 8.5592×10^{-4}. These low error values imply that the model fits the data well with minimal error. These results imply that accuracy has been maintained with the simplification of the model.

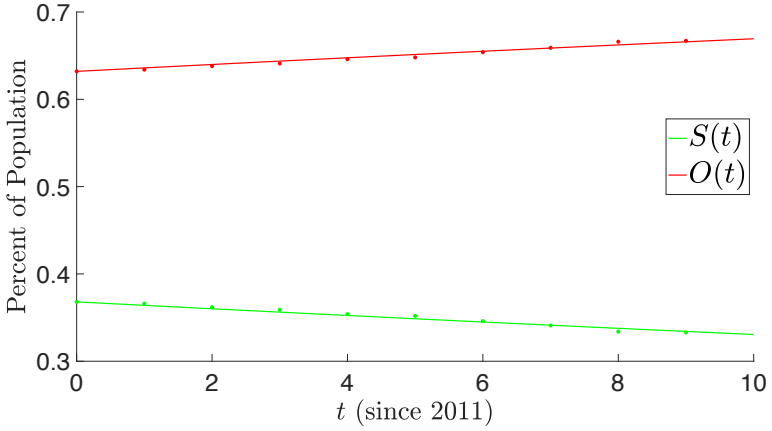


FIGURE 3.6. Solution curves for equations 3.4 and 3.5 optimized to United States data (CDC, 2023a) with a β = 0.03 and r = 0.005.

The two compartment model was optimized over both β and r and then optimized over just β keeping r constant at 0.005. The difference between these approaches was minimal as the mean squared difference for the β values was 3.7711×10^{-5} and the mean difference in the error was 1.73534×10^{-7}. Because of the similarity of the determined errors, for the remainder of the analysis, r was assumed to be equal to 0.005. Keeping r constant is consistent with the assumption that the rate at which people are able to lose weight is not dependent upon what state they live in. Additionally, r is likely on the lower end of the estimated range because the model does not take into account different obesity classifications such as morbid obesity for which the theorized r values can be as low as 1 in 1290 for men and 1 in 677 for women. (Fildes et al., 2015).

The results from optimizing β values for each state based upon data available were then plotted versus the mean percentage of overweight and obese individuals over 10 years in Figure 3.7. There was a strong inverse correlation ($R^2 = 0.911$) between these two values, meaning that states with a lower combined overweight and obese percentage had higher β values. This relationship could imply that states with higher overweight percentages could be approaching a hypothetical steady state where the limited healthy weight population is slowing down the transition to an overweight class, as there are fewer healthy weight individuals.

The β values for each state were then compared to other societal factors for each state. A correlation between various parameter values and the respective β value for each state was expected.

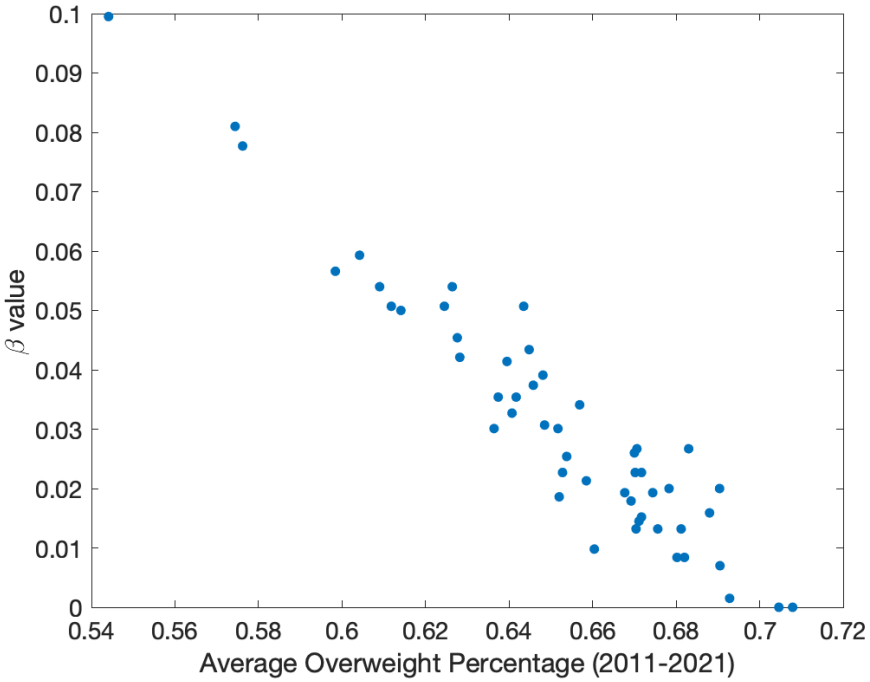


FIGURE 3.7. Comparison of optimized β values for each state versus the mean percentage of overweight and obese individuals from 2011-2021.

This correlation was expected to be of a similar strength as the correlation that existed between these societal factors and the raw data. These β values were compared to factors such as the percentage of people in poverty, the percentage of people meeting exercise standards, and the percentage of people getting sufficient sleep. Other societal influences such as mental health, food insecurity, health insurance, and average cost per meal were also analyzed but the previously mentioned factors resulted in the strongest correlations. For each of these parameters, there was a weak correlation between the parameter value and β, if there was any correlation at all. The states with higher poverty percentages have lower β values, states with higher percentages of people meeting exercise standards have higher β values, and states with more people meeting sufficient sleep standards have higher β values. However, when the β values were compared with the change in poverty levels over time for each state, the R^2 value for this comparison is lower than the R^2 value for the correlation between β and poverty levels in 2019. For the correlation between R^2 and the change in poverty levels, the R^2 value was 0.0504, whereas the correlation between β and poverty levels in 2019, the R^2 value is 0.2254. Thus, less of the variation in β values can be explained by the variation in the rate of change in poverty levels than the variation in poverty levels in 2019 alone.

A multivariate linear regression was conducted to estimate β values. When the inputs to this multivariate regression model are health insurance, exercise, food insecurity, poverty and sleep, after 1000 model runs, the average R^2 value is approximately 16%. A substantial amount of the variation in these beta values are likely not able to be accounted for with only the variation in these demographic values or these relationships are not able to be captured by linear trends.

3.3. Agent-based Model

3.3.1. Methods

Agent-based models are comprised of interacting, autonomous "agents" whose behaviors are governed by simple rules, and interactions with other agents, which in turn influence their behaviors (Macal and North, 2005). NetLogo (Wilensky, 1999) is a programmable modeling environment for simulating natural and social phenomenon. It models complex systems developing over time. Using NetLogo (Wilensky, 1999), an agent-based model of obesity was developed. This model was inspired by previously published infection models in NetLogo (Stonedahl and Wilensky, 2008). However, most of the code was developed from scratch in order to account for the differing model dynamics and the ability of nodes to die and reproduce.

The goal of the initial model was to capture the dynamics seen in the system of two differential equations utilizing an agent-based model. Thus, the population was divided into two categories, obese and then healthy-weight which in this model is inclusive of the overweight population. Ideally, this model could match the trends seen in the data well with minimal error. Eventually, the goal is to be able to differentiate the probability of becoming obese in different nodes based upon their demographic characteristics, such as income or education.

In NetLogo, each agent is referred to as a turtle. In the model, there were 1000 agents or turtles. Each turtle is created at a random x, y coordinate on the grid. A random number between 0 and 100 is assigned to each turtle. If this number is greater than the 100 minus the initial obese population percentage, than this turtle will be designated as obese to start. This platform enables us to randomly generate a population with approximately the same number of obese individuals as would be present in our population. Otherwise, the turtle will be designated as healthy weight because we are assuming that those who are not obese are in the healthy weight population. After the population is initialized, each turtle goes through a series of commands, which are governed by certain probabilities. For each time step, each turtle moves around the board with a random probability. Each turtle is then assessed to see if it will die during that time step. Again, a random number between 0 and 100 is assigned to each turtle and if that number is less than the death rate, than the turtle dies and a new turtle is "born". This designation keeps the population at a constant number of people, as was also assumed in the ODE model. This new turtle is always born to be healthy weight. Finally, the healthy weight turtles in the population are each given a random number and if this random number is less than the likelihood that a given turtle will become obese in the time step, the turtle will become obese. These commands are repeated for every time step in the simulation.

For our model of the United States population, the following assumptions were held:

- Death rate of 1% per year (CDC, 2023b).
- Birth rate of 1% per year. Birth rates were assumed to match death rates of keep population constant.
- All births are classified as health-weight.
- The rate of becoming obese for any given health-weight node was 0.56% per year which was determined based upon trends for the past ten years (CDC, 2023a).
- A population of 1000 is sufficiently large. This assumption could be analyzed through convergence testing.
- 27% of the population was assumed to be initially obese because that was the national obesity percentage in 2011 (CDC, 2023a).

However, when the agent based model was ran with these parameters, after ten years, the obesity level was only around 28%. This result does not seem to correspond to the 33% obesity rate seen in 2021, which is the year the model is trying to replicate. Thus, the probability that any healthy-weight node becomes obese was changed to 1.56%. This necessary increase is to account for the fact that each baby is born healthy-weight. After this change was made, for one run of the simulation after ten years, 35% of the population is obese, which is much closer to what the data indicates it should be. These transition percentages are displayed in a flow chart indicating the logic behind the model in Figure 3.8.

Simulations of these dynamics were run for 10 years. Each simulation resulted in a time-series that depicted the levels of obesity, overweight, and healthy weight at each of the years. The error value between the data and the model results was computed for both singular runs and the average of 5000 runs. Then, the dynamics were run 5000 times for 100 years to predict what obesity dynamics might look like in the distant future.

Next, the NetLogo model was generalized to other areas and regions of the United States. Levine (2011) studied obesity on a county level. Thus, obesity trends were forecasted with the NetLogo model on a county level. Counties in North Carolina were analyzed, more specifically Buncombe and Robeson counties. These counties were analyzed because they had the lowest and highest levels of obesity respectively in 2022 (Johnson, 2023). Robeson county has a poverty rate of 27.9% whereas Buncombe county has a poverty rate of 11.7% (U.S. Census Bureau, 2022). It was seen that Buncombe county had an obesity level of 22% in 2011 and 28% in 2022 with an average yearly rate of change of 0.55%. Robeson county had an obesity level of 39% in 2011 and 44% in 2022 with an average yearly rate of change of 0.45% (Johnson, 2023). Similar trends are seen that counties with higher levels of poverty had higher levels of obesity. However, again, those with higher rates of poverty are seeing lower average annual rates of change in obesity levels. For Buncombe county, the initial obese population was 22% and that the yearly increase in obesity levels which was calculated based upon data from the last 10 years was 0.55% which was changed to 1.55% to account for the births. For Robeson county, the initial obese population was 39% and that the yearly increase in obesity levels which was calculated based upon data from the last 10 years was 0.45% which was changed to 1.45% to account for the births (Johnson, 2023).

3.3.2. Results

Simulations were first ran in NetLogo for 10 years at a time, modeling obesity trends from 2011 to 2021. Visually, the data aligned with the agent based model well as seen in Figure 3.9. In order to quantify the accuracy of the model to the data, the mean square error was calculated. The mean square error was calculated by taking the square of the quantity which is the data obesity level for a given year minus the model estimation for that year. Then, we averaged these quantities over each year that we had data. The results of one Netlogo run was compared to the data by calculating the mean square error, which was 0.218% for one run. After 5000 NetLogo simulations, the mean square error was 2.06%. The variation between errors for subsequent runs is due to the inherent randomness of the agent-based model. These error percentages indicate that the model is a good indication of trends in the data. This error seems relatively low considering the simplicity of the model which could imply that perhaps obesity trends are not as complex as previous researchers have theorized.

When run for 100 years, the long term obesity trends are able to be forecasted as seen in Figure 3.10. These trends predicted by the model result in obesity levels leveling out to approximately

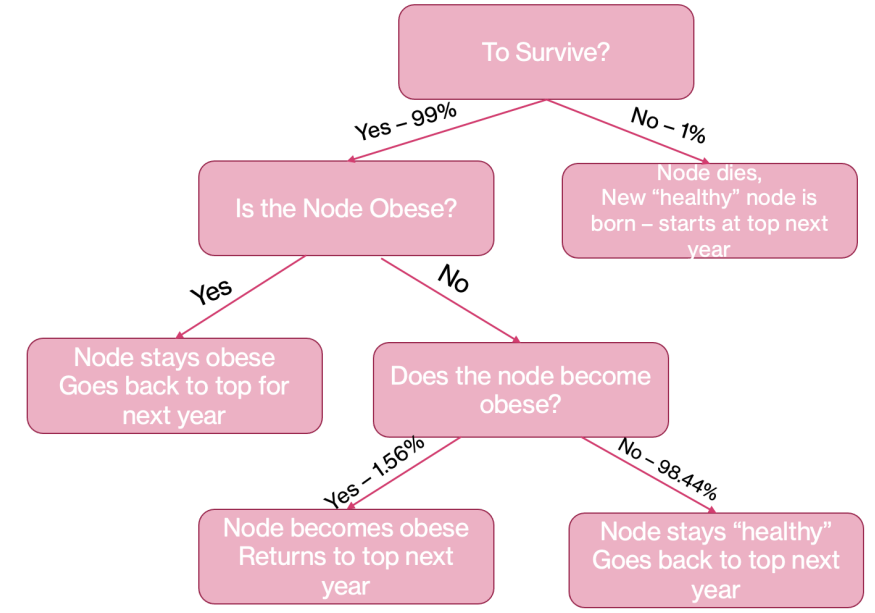


FIGURE 3.8. Flowchart of decisions made in agent based model

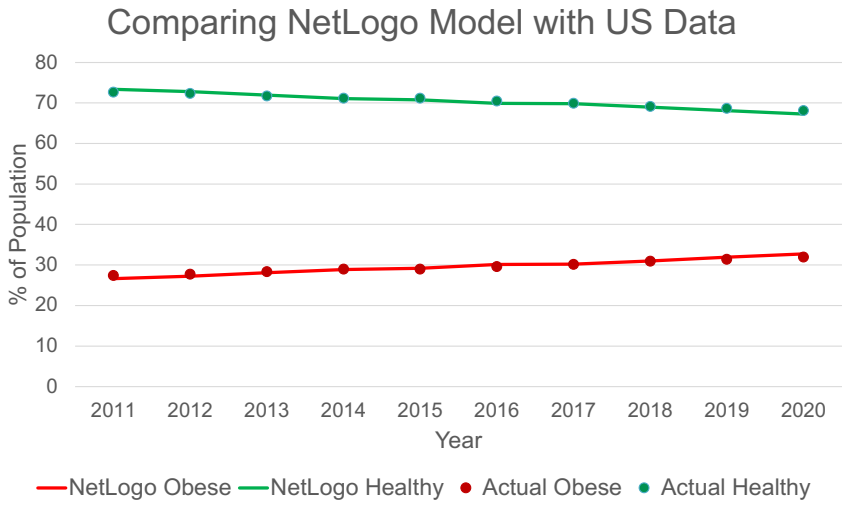


FIGURE 3.9. Comparing NetLogo Results out 10 years with data.

58% after 100 years (post 2011) after 5000 runs. This percentage only includes the obese population, not the overweight population. If the overweight population is assumed to stay consistent around 30%, that would mean that around 2090-2100, approximately 60% of the population will be obese, 30% of the population will be overweight and only 10% of the population will be healthy weight by current definitions. The predicted levels of obesity would result in an increase in related diseases that obesity puts people at an increased risk for such as Type 2 diabetes and heart disease. The increase in these diseases would cause greater strain on the healthcare system within the United States than is currently occurring.

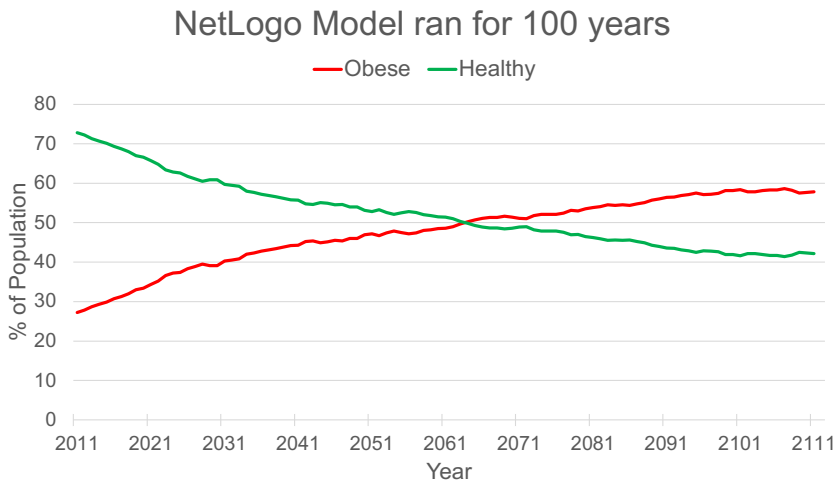
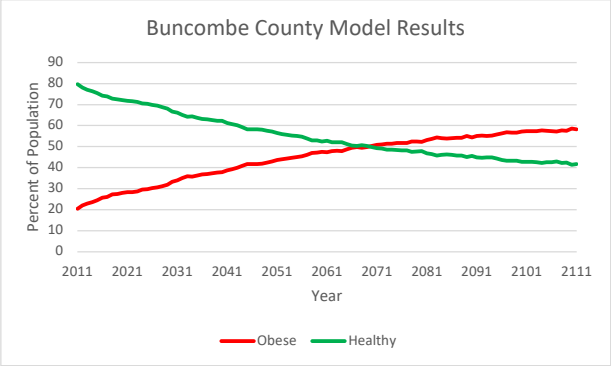
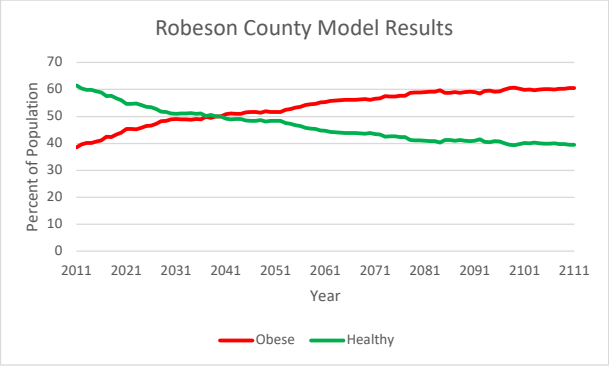


FIGURE 3.10. NetLogo predictions from 2011 for 100 years.



(A) Buncombe County predictions



(B) Robeson County Predictions

FIGURE 3.11. NetLogo predictions for Buncombe and Robeson counties.

Simulations were then ran for two different counties in North Carolina. Predictions of these counties were plotted in NetLogo as seen in Figure 3.11. Obesity levels in both of these counties eventually level out, or reach what appears to be a steady state, if a constant annual rate of change of obesity levels is assumed. Robeson obesity levels reach a steady state of slightly more than 60% and Buncombe obesity levels reach a steady state of slightly less than 60%. Robeson's point of intersection of obesity percentages and non-obesity percentages occurs about 30 years earlier.

4. Conclusion and Discussion

During this project, three different types of modeling techniques were explored. Linear regression models, an SIR-type differential equation model and agent based models were developed. These models all attempted to incorporate societal factors such as poverty into the model predictions and analysis. These models had varying levels of success.

First, a linear regression model was developed. The benefit of starting with this model is it provided a baseline of what trends could be expected while developing more complex models.

However, the issue with just looking at linear correlations is they are unable to capture the long-term trends of obesity levels within the United States. Obesity is unlikely to continue to grow at the same linear rate in the next ten years as it did in the previous ten years. Obesity trends have proven not to be linear, and they also cannot be explained by just one societal factor. Rather, obesity trends are likely explained by a combination of environmental, demographic, social and biological influences. Various aspects of weight gain and loss that are not fully understood by the medical and public health communities adds some randomnesses to who becomes obese. This weak correlation between these factors and obesity could be due to the genetic component of obesity. Bouchard (2021) suggests that the genetics account for 40% to 50% of the variability in body weight status. This influence is lower amongst normal weight individuals (30%) and substantially higher amongst individuals with obesity and severe obesity (about 60% - 80%). Perhaps, since the model does not directly take genetics into account, we can not expect to get significantly higher correlations between the social determinants of health and the obesity levels in a specific year.

Next, a differential equation model was explored. First, a three compartment SIR-type model was developed. This model was very difficult to create an accurate parametrization for as there were some parameters that seemed to have little to no impact on the model output. Then, the differential equation model was simplified to a two-compartment model that had a more logical and seemingly accurate parametrization. However, the differential equation model still had issues. One of which is that the model assumes that interaction with an overweight or obese person is necessary in order to become overweight or obese which is not based in evidence. Even Christakis and Fowler (2007) only claim that interaction, especially close interactions, change the likelihood of obesity, but does not claim that interaction is necessary for someone to become obese. Additionally, the remaining lack of sensitivity to the r parameter indicates that the differential equation model might be making the problem more complex than required. Especially because the scientific community has little to no confidence about what long term trends will look like. Since the causes of the dramatic increase in obesity levels are not well understood, many assumptions with limited factual support have to be made in order to simplify obesity trends into a model. Finally, the multivariate linear regression between β and the different societal factors indicate that perhaps there might be more at play. The trends between these parameters and the obesity data itself are significantly stronger. Perhaps instead of just looking at the rates of poverty, food insecurity, or exercise, deeper insight might be gained if the analysis took into consideration how these factors change over time with respect to obesity. However, the issue with this suggested analysis is that many of these factors are only measured once every couple of years. Over the past ten years, data is limited and conclusions are difficult to make based upon limited data. Either the relationship between the obesity model and these societal factors is not being accurately measured, the obesity model is not properly taking into account the various societal factors at play or the societal factors have little impact on population obesity models. In order to analyze these factors properly, the model must be adapted so that these factors influence the obesity model projections.

Finally, an agent-based model was created to model obesity levels. The benefit of the agent-based model over the differential equation model is that it was able to accurately portray the trends in the available data with a relatively simple model. This accuracy is reflected by the relatively small error seen in the results of the model compared to the data. However, the strengths of agent-based modeling could be additionally taken advantage of through a further investigation of individual interactions and their influence on obesity trends. One possibility is to network nodes to determine the social influence of other nodes. However, this possibility becomes difficult to do in the current software due to nodes constantly being deleted and regenerated throughout the

time span of the model to reflect deaths and births respectively. Additionally, the probabilities of change for each node could be governed by different environmental factors, with each region of the model representing different environments. These environments could take into consideration the difference between living in a poverty-stricken, rural, food desert, and living in a wealthy, suburban neighborhood with easy access to healthy foods and opportunities for exercise. The preliminary investigation into agent based modeling shows the potential of agent based modeling further expanding the knowledge surrounding the dynamics of obesity trends. The agent based model also supports the idea that perhaps poverty and different societal factors just change the starting point, but the overall trends between locations are very similar or that higher levels of obesity have lower transition rates. This idea of delayed trends could be investigated further with more access to historical data.

One thing of note that was discovered during the investigation into North Carolina counties was that Wake County had very unusual obesity trends within the state. Wake County's obesity level was 27% in 2011 and 28% in 2022. Wake County has been able to keep their obesity levels relatively constant while most other places have had increases of approximately 10% in the last ten years. Perhaps this is indicative of prevention strategies working or other larger factors at play, such as gentrification.

The linear regression models did not seem sufficient enough to predict obesity trends into the future since obesity trends have proven not to follow a linear trend. However, the linear regressions gave insight into what demographic factors likely have the greatest impact on obesity levels, and enabled the focus to be narrowed down to poverty for future investigations. An SIR based model captured data well, but a complex model is not needed to establish trends. The 2-equation system of ODEs appeared to fit data as well as the 3-equation system. Additionally, when the ODE model was simplified down to two equations, it was able to be more reflective of population dynamics. However, the ODE model is limited in its predictive potential because the β values seem relatively arbitrary when compared to population demographics. Finally, the agent-based model seemed to be able to capture the population dynamics accurately with a relatively simple probabilistic model.

A potential path of future research would be to investigate how different intervention strategies influence obesity trends. Various intervention strategies that could be considered include weight loss drugs, healthy eating initiatives, exercise initiatives, and poverty reduction strategies. Most notably, it would be interesting to conduct a cost-benefit analysis of the newly approved GLP-1 class of weight loss drugs that include Wegovy and Ozempic. This research could include an investigation of what percentage of the obese population would need to take these drugs in order to substantially reduce the related medical and other related costs of obesity. Additionally, the overall cost of these drugs could be taken into consideration, especially because they are currently theorized to have to be taken indefinitely and they are currently advertised to have a monthly cost of greater than $1,000. Another source of future research could be to investigate the impact of reductions in obesity levels on medical costs and the rates of other diagnosable diseases that obesity influences, such as type 2 diabetes and heart disease.

Acknowledgement

The authors would like to thank the reviewers. The authors would also like to thank the financial support of the Honors Program and the Undergraduate Research Program at Elon University.

References

Bouchard, C. (2021). Genetics of obesity: what we have learned over decades of research. *Obesity*, 29(5):802–820.

CDC (2014). CDC - About BRFSS. https://www.cdc.gov/brfss/about/index.htm. Accessed: 2024-02-18.

CDC (2022a). Adult obesity facts. https://www.cdc.gov/obesity/data/adult.html. Accessed: 2024-02-13.

CDC (2022b). Defining adult overweight & obesity. https://www.cdc.gov/obesity/basics/adult-defining.html. Accessed: 2024-02-13.

CDC (2023a). Data, trend and maps. https://www.cdc.gov/nccdphp/dnpao/data-trends-maps/index.html. Accessed: 2024-02-13.

CDC (2023b). FastStats - Deaths and Mortality. https://www.cdc.gov/nchs/fastats/deaths.htm. Accessed: 2024-02-13.

CDC (2024). Chronic Disease Indicators (CDI) Data [online]. https://nccd.cdc.gov/cdi. Accessed: 2024-02-18.

Christakis, N. A. and Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4):370–379.

Delavani, Aldila, D., and Handari, B. D. (2021). Effect of healthy life campaigns on controlling obesity transmission: A mathematical study. *Journal of Physics: Conference Series*, 1747(1):012003.

Dinour, L. M., Bergen, D., and Yeh, M.-C. (2007). The food insecurity–obesity paradox: a review of the literature and the role food stamps may play. *Journal of the American Dietetic Association*, 107(11):1952–1961.

Fildes, A., Charlton, J., Rudisill, C., Littlejohns, P., Prevost, A. T., and Gulliford, M. C. (2015). Probability of an obese person attaining normal body weight: cohort study using electronic health records. *American journal of public health*, 105(9):e54–e59.

Fryar, C. D., Carroll, M. D., Ogden, C. L., et al. (2012). Prevalence of overweight, obesity, and extreme obesity among adults: United states, trends 1960–1962 through 2009–2010. *Hyattsville, MD: National Center for Health Statistics*.

Hales, C. M., Carroll, M. D., Fryar, C. D., and Ogden, C. L. (2020). Prevalence of obesity and severe obesity among adults: United states, 2017–2018. *NCHS Data Brief, no 360. Hyattsville, MD: National Center for Health Statistics*.

Hernandez, B. L. M., Reena, I., Strickland, G., and Ruiz, C. (2015). The impact of infant feeding method on childhood obesity/overweight levels of children at ages 2, 3, and 4 years. *Global Journal of Health Education and Promotion*, 16(2).

Johnson, S. (2023). County Health Rankings & Roadmaps - Adult Obesity. https://www.countyhealthrankings.org/explore-health-rankings/county-health-rankings-model/health-factors/health-behaviors/diet-and-exercise/adult-obesity. Acccessed: 2024-02-18.

Kompaniyets, L., Goodman, A. B., Belay, B., Freedman, D. S., Sucosky, M. S., Lange, S. J., Gundlapalli, A. V., Boehmer, T. K., and Blanck, H. M. (2021). Body mass index and risk for covid-19–related hospitalization, intensive care unit admission, invasive mechanical ventilation, and death—united states, march–december 2020. *Morbidity and Mortality Weekly Report*, 70(10):355.

Levine, J. A. (2011). Poverty and obesity in the us. *Diabetes*, 60(11):2667.

Macal, C. and North, M. (2005). Tutorial on agent-based modeling and simulation. In *Proceedings of the Winter Simulation Conference, 2005.*, pages 14 pp.–.

Ogden, C. L., Lamb, M. M., Carroll, M. D., and Flegal, K. M. (2010). Obesity and socioeconomic status in adults: United states 1988–1994 and 2005–2008. *NCHS data brief*, 50:1–8.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Santonja, F.-J., Villanueva, R.-J., Jódar, L., and González-Parra, G. (2010). Mathematical modelling of social obesity epidemic in the region of valencia, spain. *Mathematical and Computer Modelling of Dynamical Systems*, 16(1):23–34.

Smith, D., Moore, L., et al. (2004). The sir model for spread of disease-the differential equation model. *Convergence*.

Smith, N. R., Zivich, P. N., and Frerichs, L. (2020). Social influences on obesity: Current knowledge, emerging methods, and directions for future research and practice. *Current nutrition reports*, 9:31–41.

Stonedahl, F. and Wilensky, U. (2008). NetLogo virus on a network model. *Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.*

Thomas, D. M., Weedermann, M., Fuemmeler, B. F., Martin, C. K., Dhurandhar, N. V., Bredlau, C., Heymsfield, S. B., Ravussin, E., and Bouchard, C. (2014). Dynamic model predicting overweight, obesity, and extreme obesity prevalence trends. *Obesity*, 22(2):590–597.

U.S. Census Bureau (2022). U.S. Census Bureau QuickFacts: North Carolina. `https://www.census.gov/quickfacts/fact/table/wakecountynorthcarolina,robesoncountynorthcarolina,buncombecountynorthcarolina,NC/IPE120221`.

USDA (2023). USDA ERS - Key Statistics & Graphics. `https://www.ers.usda.gov/topics/food-nutrition-assistance/food-security-in-the-u-s/key-statistics-graphics/#map`. Accessed: 2023-10-26.

Ward, Z. J., Bleich, S. N., Long, M. W., and Gortmaker, S. L. (2021). Association of body mass index with health care expenditures in the united states by age and sex. *PloS one*, 16(3):e0247307.

Wilensky, U. (1999). NetLogo. *Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.*

(D. DaSilva) Department of Mathematics and Statistics, Elon University, Elon, NC 27244, USA
Email address, Corresponding author: `ddasilva@elon.edu`

(K. Yokley) Department of Mathematics and Statistics, Elon University, Elon, NC 27244, USA
Email address: `kyokley@elon.edu`