The Difference of Income Levels and Their Influence on the Percentage of Obese Adults in America from 2011-2022

Alison Miller

5/1/2024

Obesity is a costly, deadly disease that affects many Americans. The Centers for Disease Control and Prevention (CDC) defines obesity as having a Body Mass Index (BMI) of 30 or higher. Obese adults are at an increased risk of suffering from chronic conditions such as high blood pressure, type 2 diabetes, breathing, and joint problems, along with psychological issues like anxiety and depression. Specific demographics are affected by obesity more than others. Middle-aged, 40-59 years old adults are more likely to be obese than adults aged 20-39 years or adults 60 and older. Non-Hispanic Black adults have the highest age-adjusted rate of obesity, followed by Hispanic, non-Hispanic White, and non-Hispanic Asian adults. Adults with less education are more likely to be obese than those with a college degree (Centers for Disease Control and Prevention [CDC] 2023). Income level is another way to categorize people with obesity. Removing all other factors such as race, education, and age, it was theorized that there is a greater percentage of obese adults with a lower income than those with a higher income level.

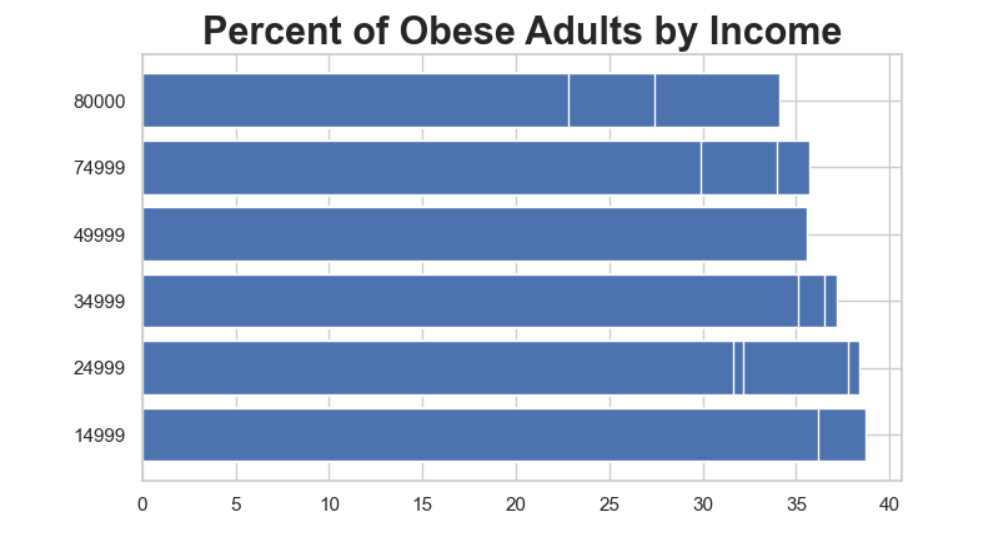
The dataset selected for the study was the Nutrition, Physical Activity, and Obesity from the Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS is a telephone survey conducted by the CDC that collects health-related data about residents in the United States. The survey asks questions about chronic health conditions, health-related risk behaviors, and the use of preventive services. Over 400,000 adults are asked a set of core questions each year. The data is weighted using iterative proportional fitting, reducing the potential for bias (CDC, 2014). Iterative proportional fitting (IPF) creates two-dimensional tables by combining two one-dimensional tables with respect to their given totals. This is useful when combining data to create a matrix. A matrix can be made using race, income, and education levels while still respecting the row and column totals (Iterative Proportional, 2024).

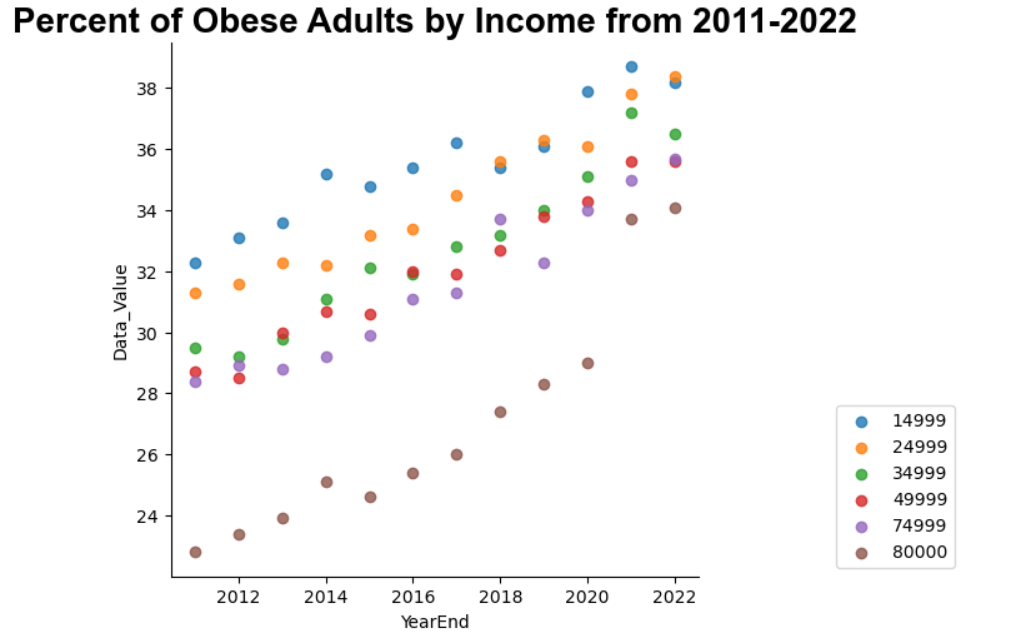
The Nutrition, Physical Activity, and Obesity dataset includes information such as the years the data was collected, location, questions asked, sample size, age of the person the information was collected from, education, gender, income, and race/ethnicity. The dataset can be filtered to select the focus variable, such as certain states, years, or particular questions asked. The focus of this study was the percentage of adults aged 18 years and older who have obesity and what income level they reported. Data from the entire United States was used instead of focusing on a specific state or geographic area. The target variable was the percentage of adults aged 18 years and older who have obesity. Education, race/ethnicity, gender, and individual age groups were removed from the dataset. Hence, income level was the only independent variable left.

Next, the data was cleaned by removing unnecessary columns: race/ethnicity, education, gender, high and low confidence limits, class ID, topic ID, sample size, age, year start, data source, class, topic, data value unit, data value type, data value alt, data value footnote symbol, data value footnote, total, geolocation, data value type ID, and location ID. Year start was dropped because the year start and the year-end were within the same year; therefore, only one category was necessary. All data not reported values were dropped. The data was also cleaned to remove commas, dollar signs, and words. The income categories were changed from a range to the maximum value of the income range. The category Less than $15,000 was changed to 14999. The income category of $75,000 or greater was changed to 80000. Data cleaning removes potential errors and biases. Categories with non-reported data can skew the results. The income range was eliminated to make the income category consistent (Bhandari, 2023b).

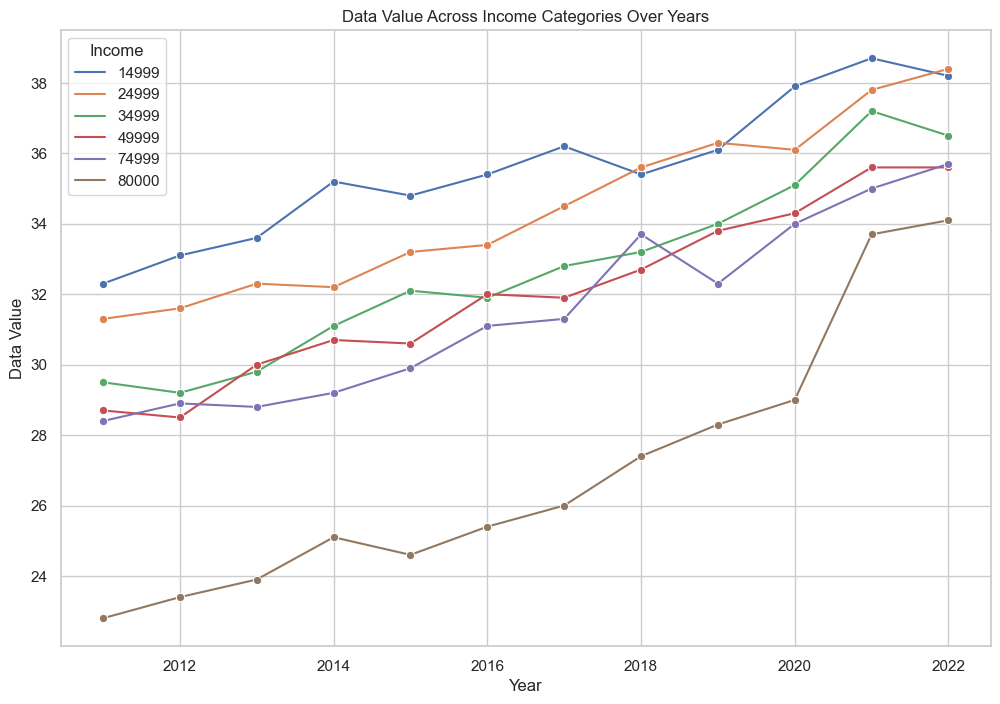
Descriptive statistics were used to understand the data, encapsulating its characteristics (Bhandari, 2023a). The mean and median were calculated across all income levels to determine if the data was skewed. The mean measures the average of a dataset, while the median represents the middle of a dataset. The mean percentage of obese adults in all income levels was 32.2%, and the median was 32.5%. The mean and the median are very close together, which indicates a near-zero skew or bias (Turney, 2023). Knowing the data is not skewed means the data has a normal standard distribution, meaning most of the data will be within two standard deviations of the mean. Since the data follows the assumption of normality and homoscedasticity, parametric tests can be performed (Kim, Tae Kim, & Park, Jae Hong, 2019). Homoscedasticity means the size of the error in our hypothesis does not change significantly within the range of our explanatory variable or income (Bevans, 2023.)

A bar graph was created to show the relationship between the percentage of obese adults and income levels. The percentage of obese adults is plotted along the x-axis, and the income levels are plotted along the y-axis. The bar length visually confirms that the median and mean calculation was correct (Bar Graph, n.d.).

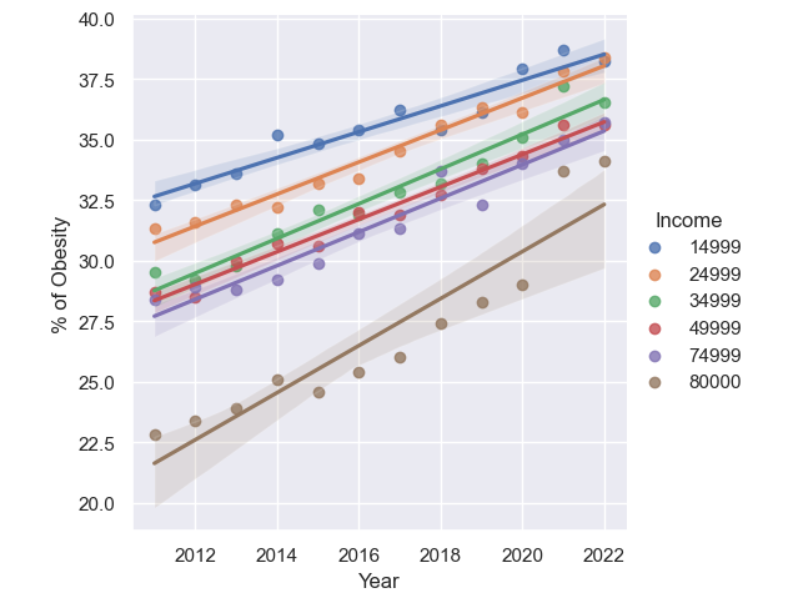


The following graphs created were scatterplot graphs. Scatterplot graphs visually identify a relationship between the changes observed in variables (Minnesota Department of Health, 2023). The percent of obese adults is plotted along the y-axis, and the year is plotted along the x-axis. The income levels were shown using different color hues. Each color represents a different income level. This graph demonstrates the percentage of obese adults through the years categorized by income level. An income of $14,999, shown in blue, represents the highest proportion of obese adults. The income level of $80,000, shown in brown, is at the bottom of the graph. As income increases, the percentage of obese adults decreases. This proves the theory that there is a more significant percentage of obese adults with a lower income than those with a higher income level. Another interesting observation is the increase in the percentage of obese adults as the years increase. Every income level experiences an increase as time passes.

The data from the scatterplot overlaps at times; therefore, a line graph was also created to show the increase in the percentage of obese adults at different income levels over the years. Line graphs are a valuable way to visualize trends over a period of time. Again, the percentage of obese adults was plotted along the y-axis, the year was plotted along the x-axis, and the income levels were depicted by color. The sharp increase in obesity in the income level of $80,000 between 2020 and 2021 is better illustrated using the line graph. Also, there was a large gap between all of the income levels and the $80,000 income level until 2021, when the gap drastically shrank.

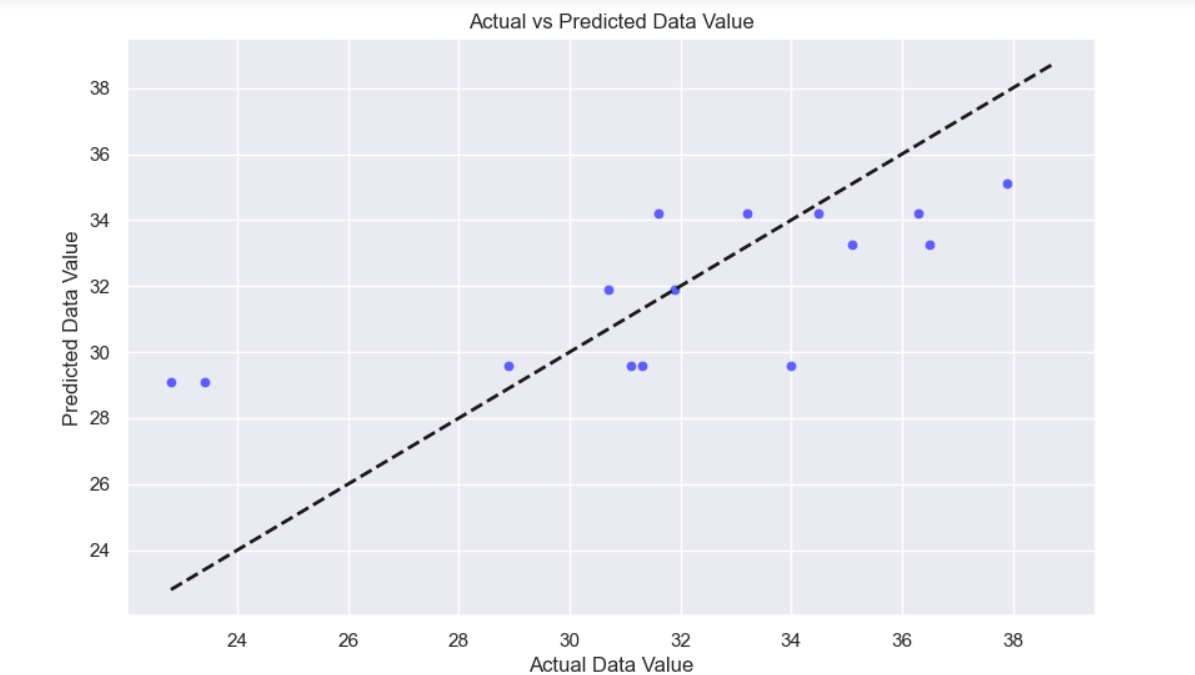


A multiple linear regression graph is another way to describe the relationship between variables. Regression models view the relationship by fitting a line between the observed data. The percentage of obese adults is plotted on the y-axis, and the year is plotted on the x-axis. Income is represented by different color hues and is used to differentiate between the income groups. The slope of our regression line is the equation y = mx1 +mx2 + mx3 + +b. Obesity = high-income level + mid-income level+ low-income level + b. The points tightly align with the slope in a positive direction. Obesity rates did not change much within the same income category, except for the category $80,000. In 2021, the percentage of obese adults dramatically increased compared to previous years, which is shown by how far away they are from the best-fitted line.



After a relationship between the percentage of obese adults and income was visualized, a machine learning model was developed. Machine learning, a subset of Artificial Intelligence (AI), uses data and algorithms to copy situations to analyze, predict, and study behaviors and events. Data was split into training and test models to create a machine learning model. The training data trains the model by teaching an algorithm to reach an expected outcome. 80% of the data was allocated to the training model, and 20% was used to test the training model. The test model checks whether the training algorithm is correct (Smolic, 2022). A mean squared error was calculated to check the training model. The mean square error measures the average squared difference between the predicted and actual target values (Mean Square Error, 2023). The mean squared error of the testing model was 8.86. A small mean square error proves the model’s predictions are closer to the target values.

Another statistic to evaluate the model is R2. In the article” How to Interpret R-squared in Regression Analysis,” R2 is defined as “measuring the strength of the relationship between your model and the dependent variable.” R2 assesses the scatter of points around the fitted regression line. The closer the R2 value is to 100%, the more the model explains the variations in the response variable around the mean. A large R2 value means the data will be plotted closer to the regression line, which means your training and test models are accurate. The R2 value for the test model was 50%, which means our training and test model need improvements. A 50% R2 value indicates that about 50% of the variance in the percentage of obese adults can be explained by the independent income variables. As seen in the graph below, some of the data is close to the regression line, but there are outliers.



Lastly, the slope and the intercept of the model were calculated. The slope of a regression line means the change of the dependent variable (percent of obese adults) increases or decreases the independent variable (income levels). The intercept represents the value of the dependent variable when the independent is zero. The slope and intercept of the linear regression model were -9.2 and 36.5, respectively. For an income level of zero, the expected percentage of obese adults would be 36.5%. As the income increases, the percentage of obese adults would decrease by 9.2 percentage points. Therefore, with an income of $15,000, the expected percentage of obese adults would be 27.3%. This is not depicted in the graph above; the percentage of obese adults does not drop by 9.2 points. The training model is inaccurate because other factors besides income affect the percentage of obese adults.

Obesity is a rising epidemic in America. Obese individuals have higher rates of chronic conditions, increased healing time, and poorer surgical outcomes. Additionally, they report higher levels of depression, lower self-esteem, and dissatisfaction with life. The dataset selected for the study was the Nutrition, Physical Activity, and Obesity from the Behavioral Risk Factor Surveillance System (BRFSS). Information on the percentage of obese adults in America was analyzed in conjunction with reported income levels to determine if there was a correlation between lower income levels and a higher percentage of obese adults. This theory was proven correct through statistical and graphical means. Understanding the factors contributing to obesity can allow health officials to allocate more aid and support to those patients. Patients in lower income brackets can be given dietary and physical fitness advice, extra monitoring, and other considerations to help prevent obesity. Patients and healthcare providers can work together to stop obesity before it becomes a deadly health condition.

References:

Bar Graph. BYJU’S. (n.d.). https://byjus.com/maths/bar-graph/#:~:text=The%20bar%20graph%20helps%20to,changes%20in%20data%20over%20time.

Bevans, Rebecca. Simple Linear Regression | An Easy Introduction & Examples. Scribbr. June 22, 2023a. https://www.scribbr.com/statistics/multiple-linear-regression/.

Bhandari, Pritha. Descriptive Statistics | Definitions, Types, Examples. Scribbr. June 21, 2023a. https://www.scribbr.com/statistics/descriptive-statistics/.

Bhandari, Pritha. What is Data Cleansing? Definition, Guide & Examples. Scribbr. June 21, 2023b. https://www.scribbr.com/methodology/data-cleansing/.

Centers for Disease Control and Prevention. (September 21, 2023). Adult Obesity Facts*.* https://www.cdc.gov/obesity/data/adult.html.

Centers for Disease Control and Prevention. (May 16, 2014). Behavioral Risk Factor Surveillance System. https://www.cdc.gov/brfss/about/index.htm.

## Centers for Disease Control and Prevention. (December 7, 2023). Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System. https://data.cdc.gov/Nutrition-Physical-Activity-and-Obesity/Nutrition-Physical-Activity-and-Obesity-Behavioral/hn4x-zwk7/about\_data.

Frost, Jim. How to Interpret R-squared in Regression Analysis. Statistics by Jim. (n.d.) https://statisticsbyjim.com/regression/interpret-r-squared-regression/.

Iterative Proportional Fitting. Travel Forecasting Resource. January 25, 2024. https://tfresource.org/topics/Iterative\_Proportional\_Fitting.html.

Kim, Tae Kim, & Park, Jae Hong (2019). More about the basic assumptions of t-test: normality and sample size. Korean Journal of Anesthesiology, *72*(4), 331–335. https://doi.org/10.4097/kja.d.18.00292.

Mean Square Error (MSE). Encord. 2023. https://encord.com/glossary/mean-square-error-mse/#:~:text=In%20the%20fields%20of%20regression,target%20values%20within%20a%20dataset.

Scatterplot. Minnesota Department of Health. May 01, 2023. https://www.health.state.mn.us/communities/practice/resources/phqitoolbox/scatterplot.html#:~:text=A%20scatter%20plot%20identifies%20a,a%20relationship%20between%20two%20variables.

Smolic, Hrvoje. Training Data vs Test Data in Machine Learning – Essential Guide. LinkedIn. September 12, 2022. https://www.linkedin.com/pulse/training-data-vs-test-machine-learning-essential-guide-smolic-/.

Turney, Shaun. Skewness | Definition, Examples & Formula. Scribbr. November 10, 2023. https://www.scribbr.com/statistics/skewness/.