

The United States of Obesity

Matt Malloure
Grand Valley State University
matt.malloure@gmail.com

Mary Richardson
Grand Valley State University
richamar@gvsu.edu



Diann Reischman
Grand Valley State University
reischmd@gvsu.edu

Paul Stephenson
Grand Valley State University
stephenp@gvsu.edu

Published: January 2013

Overview of Lesson Plan

In this activity students will be introduced to a choropleth map and learn how such a map can be used to display data. Using the data from a choropleth map along with additional poverty data, students will explore the relationship between obesity and poverty within the United States. The students will gain experience creating a scatter plot by hand and then interpreting the strength of the linear relationship through the correlation coefficient. Next, students will use their calculator to fit a simple linear regression equation with poverty predicting obesity. Finally, the students will explore the residuals for the regression equation and interpret the results from the activity in the context of the problem.

GAISE Components

This investigation follows the four components of statistical problem solving put forth in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report*. The four components are: formulate a question, design and implement a plan to collect data, analyze the data by measures and graphs, and interpret the results in the context of the original question. This is a GAISE Level C activity.

Common Core State Standards for Mathematical Practice

1. Make sense of problems and persevere in solving them.
2. Reason abstractly and quantitatively.
3. Construct viable arguments and critique the reasoning of others.
4. Model with mathematics.
5. Use appropriate tools strategically.
6. Attend to precision.

Common Core State Standard Grade Level Content (High School)

S-ID. 1. Represent data with plots on the real number line (dot plots, histograms, and box plots).

S-ID. 6. Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

S-ID. 6a. Fit a function to the data; use functions fitted to the data to solve problems in the context of the data. *Use given functions or choose a function suggested by the context. Emphasize linear, quadratic, and exponential models.*

S-ID. 6b. Informally assess the fit of a function by plotting and analyzing the residuals.

- S-ID. 6c. Fit a linear function for a scatter plot that suggests a linear association.
- S-ID. 7. Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.
- S-ID. 8. Compute (using technology) and interpret the correlation coefficient of a linear fit.
- S-ID. 9. Distinguish between correlation and causation.
- S-IC. 3. Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

NCTM Principles and Standards for School mathematics

Data Analysis and Probability Standards for Grades 9-12

Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them:

- understand the differences among various kinds of studies and which types of inferences can legitimately be drawn from each;
- understand the meaning of measurement data and categorical data, of univariate and bivariate data, and of the term variable;
- understand histograms, parallel box plots, and scatterplots and use them to display data.

Select and use appropriate statistical methods to analyze data

- for univariate measurement data, be able to display the distribution, describe its shape, and select and calculate summary statistics;
- for bivariate measurement data, be able to display a scatterplot, describe its shape, and determine regression coefficients, regression equations, and correlation coefficients using technological tools;
- identify trends in bivariate data and find functions that model the data or transform the data so that they can be modeled.

Prerequisites

For this activity students should be practiced in summarizing and analyzing univariate data. Students should also have a basic understanding of concepts related to simple linear regression such as: scatterplots, best-fit lines, residuals, and correlation.

Time Required

In order for this activity to be completed in the classroom, two class periods will be needed. However, if a few questions are to be assigned outside of class, then only one class period is required.

Materials Required

Students will only need to bring a graphing calculator and a pencil. The instructor will provide the Activity Worksheet which will contain the data required.

Instructional Lesson Plan

The GAISE Statistical Problem-Solving Procedure

I. Formulate Question(s)

Before delving into the activity, ask students how they would go about representing obesity data for each of the 50 states. This will start a discussion about various methods that the students might have already learned before introducing them to thematic maps. Explain to the students

that a thematic map is most often used as a media outlet to show themes or topics on a map. So the map of the United States and the obesity percentage for each state would surely be an example of a thematic map. These maps summarize numerous geographical distributions of data within set regions. A choropleth map is a specific example of a thematic map, in that a summary statistic is gathered for specific areas and then each area is shaded in a color representing the magnitude of the statistic. In this activity, each state is an area and the color scale goes from green to red with the darkest green representing the lowest percentage and the darkest red representing the highest percentage. Stress to students the importance of using a statistic like the mean, median, or percentage instead of a raw count for a choropleth map. If raw counts are used, then the population in each state will essentially drive the colors instead of the real information of interest. With a better understanding of the choropleth map, explain to the students the goal of the activity.

Two of the most common concerns in the United States of late are poverty and obesity. Data have been collected on the 3-year, 2008 through 2010, average poverty rate and obesity percentage in all 50 states and D.C. In this activity students determine if the poverty rate is related to the obesity percentage within each state. More specifically, can the poverty rate be used to predict the obesity rate in a given state?

II. Design and Implement a Plan to Collect the Data

The data for this activity have already been gathered and are presented in the table on the last page of the Activity Worksheet. A copy of the data table appears below.

State	Poverty Rate	Obesity Percent	State	Poverty Rate	Obesity Percent
Alabama	16.1	32.3	Missouri	14.6	30.3
Alaska	10.8	25.9	Montana	13.4	23.8
Arizona	19.2	25.4	Nebraska	10.2	27.6
Arkansas	16.5	30.6	Nevada	13.4	25
California	15.4	24.8	New Hampshire	7.1	25.6
Colorado	11.9	19.8	New Jersey	9.8	24.1
Connecticut	8.3	21.8	New Mexico	19.1	25.6
Delaware	11.3	28	New York	15.3	24.7
District of Columbia	18.1	21.7	North Carolina	16.1	29.4
Florida	14.6	26.1	North Dakota	11.6	28
Georgia	17.5	28.6	Ohio	14.1	29.6
Hawaii	11.5	23	Oklahoma	14.3	31.4
Idaho	13.3	25.7	Oregon	12.7	25.4
Illinois	13.2	27.6	Pennsylvania	11.4	28.5
Indiana	15.6	29	Rhode Island	13.1	24.3
Iowa	10.2	28.1	South Carolina	14.9	30.9
Kansas	13.6	29	South Dakota	13.5	28.7
Kentucky	17.3	31.5	Tennessee	16.1	31.9
Louisiana	18	31.5	Texas	17.2	30
Maine	12	26.6	Utah	9.1	23.4
Maryland	9.7	27.1	Vermont	9.7	23.5
Massachusetts	10.9	22.3	Virginia	10.6	25.9
Michigan	14.2	30.5	Washington	11.2	26.4
Minnesota	10.5	25.3	West Virginia	15.7	32.2
Mississippi	21.3	34.4	Wisconsin	10.2	27.4
			Wyoming	9.6	25.4

For each state, the obesity percentage was taken from the choropleth map in the Activity Worksheet from 2008-2010 (<http://calorielab.com/news/2011/06/30/fattest-states-2011/>). As for the poverty rate for each state, that was taken from the U.S. Census Bureau (<http://www.irp.wisc.edu/faqs/faq3/table1.htm>). Even though the data don't need to be collected, there are still some important lessons for students to understand in this activity. First, ask the students why this is an observational study and not an experimental study. They should point out that the data are simply observed, meaning that the number of obese adults were counted in each state and then divided by the state's adult population. There were no interventions imposed on each subject after randomizing a subject to a treatment, etc. Secondly, since this is an observational study with the entire population of the United States, the students should realize that statistical inference is not appropriate, since the data represent a population. There was no random sample of states taken from all 50 plus D.C. so there is no need to perform inference since all of the information is present.

III. Analyze the Data

The first step in the data analysis portion of the activity is to have the students construct a scatterplot with the poverty rate on the x -axis and the obesity percentage on the y -axis. In the final plot, there should be 51 points on the plot representing the 50 states and D.C. with all axes labeled appropriately. The students are given a grid on which to neatly create this scatterplot. Figure 1 shows an example scatterplot for the data and the students' plots should be very similar.

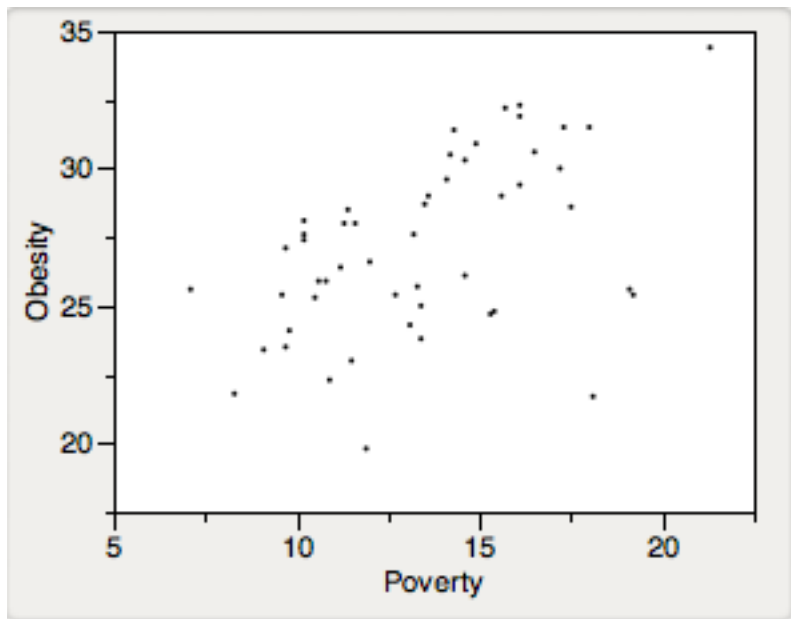


Figure 1. Scatterplot of poverty rate vs. obesity percentage.

Once the plot is finished, the students are asked to comment on the overall strength of the relationship between poverty rate and obesity percentage. Students should comment on three aspects of this plot: strength, direction, and form (linear, exponential, polynomial, etc.). In this scenario, the relationship between poverty rate and obesity percentage appears to be of medium strength, positive, and linear.

To gather a more quantitative value of this relationship, the students need to next calculate the correlation coefficient using their calculators. They need to enter the poverty rates into one list

and the obesity rates in another and then perform a linear regression. The correlation coefficient is $r = .501$. Since the students have already been given a lecture about simple linear regression concepts, they know the possible values of r are between -1 and 1 and the further from 0, the stronger the linear relationship. So when the students interpret .501 in the context of the problem, it should be similar to what they said when they interpreted the scatterplot. There is a medium strength, positive, linear relationship between poverty and obesity from 2008 to 2010.

Students are asked to use the calculator to obtain the estimated regression equation needed in question 4 on the Activity Worksheet. The estimated regression equation for this situation is

$$\widehat{obesity} = .510(poverty) + 20.323.$$

After the students plot this equation on their scatterplot, it should resemble the plot in Figure 2.

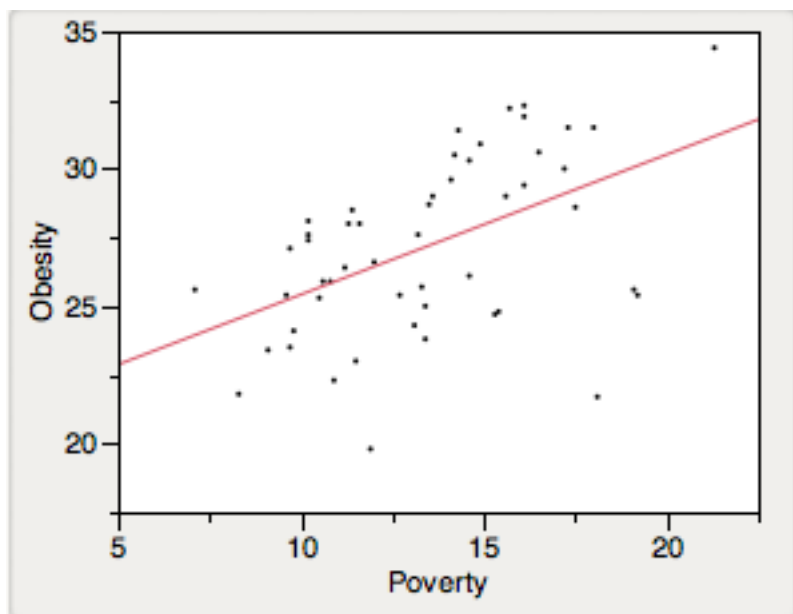


Figure 2. Poverty vs Obesity scatterplot with best-fit line added.

From this estimated regression equation, the students need to interpret both the slope and intercept terms in the context of the problem. For question 5, the slope interpretation should say something along the lines of, “for every increase of 1 percentage point in the poverty rate, we can expect the obesity percentage to increase by .510 percentage points.” As for the interpretation of the intercept term, students should understand that it corresponds to the case where the poverty rate is 0, so the interpretation should read, “when the poverty rate is 0, we expect the obesity percentage to be 20.323.” Even though the students are asked to interpret this value as an exercise in the Activity Worksheet, they need to understand that it shouldn’t be interpreted in a real-life example. There are no data points even close to a 0 percent poverty rate, so there are no values supporting the claim that at this point the expected obesity percentage is near 20. Also, in reality, there will never be a state with a poverty rate of 0, so in this regression equation, the y -intercept is simply present to improve the fit of the regression equation.

Question 7 asks students to look at their scatterplot and try to determine if any states appear to be pulling the regression line in a certain direction, hinting at the presence of outliers. From the

scatter plot it does appear that the line is pulled toward the horizontal axis making it appear not to fit the main mass of data points. Therefore, students should conclude that the two states with the lowest obesity percentages, Colorado and the District of Columbia, are outliers.

Now students are asked to find the residuals on their calculator. They first must find the predicted value for each state using the estimated regression equation and subtract the predicted value from the observed value to find the residual. The students need to take these residuals and create a histogram and stem plot by hand in question 8. Figure 3 displays an example histogram and stem plot for the residuals.

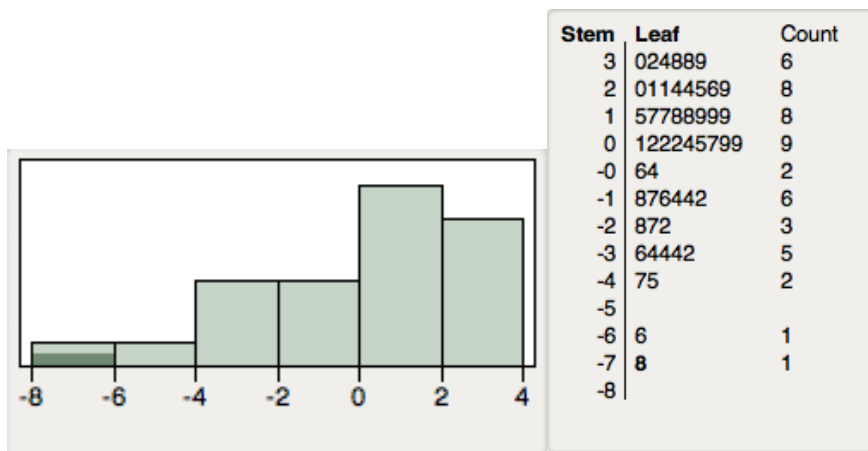


Figure 3. Histogram and stem plot of regression residuals.

Now that students have looked at outliers graphically and have calculated the residuals, they can investigate the connection between outliers and large residual values. In question 7, the students concluded that some outliers might exist and from the stem plot, there appear to be 2 residuals that are separated from the rest with values of -6.6 and -7.8 . This means that the respective observed values are 6.6 and 7.8 percentage points less than their respective predicted values. The further an observed value is from the predicted value, the larger the magnitude of the residual and thus the point might be an outlier. Notice that these extreme residuals are Colorado and the District of Columbia, the same outliers in question 7. Questions 7 through 9 allow students to practice qualitative and quantitative determination of potential outliers in regression.

One more important quantity in simple linear regression is the sum of squares for error or SSE. The goal of simple linear regression is to minimize this quantity, and the estimated regression equation is the model that indeed has the smallest SSE. The activity asks students to calculate SSE using their calculator and to do this they simply need to square the residuals and then find the sum. Explain to the students that the sum of the residuals (not squared) will equal 0 for all regression equations simply due to the construction of the regression equation. This is where the need for a sum of squared residuals comes from. The best-fit line will have the smallest SSE value because squared residuals are all positive numbers. The SSE for this regression equation is 385.751. If the students were to fit a line that they drew on the scatter plot as a best-fit line, they should expect the SSE to be greater than or equal to the SSE from the estimated regression equation's SSE. That is the motivation for question 11 in the activity. Without knowing the equation of the line, there is a chance (although slim) that a student may draw the precise regression equation found in question 4. This student would then find the same SSE, but every

other best-fit line will produce a larger SSE. So students should never suggest that the SSE of any other regression line will be less than 385.751.

In regression, extrapolation of the regression equation to points outside the domain of the independent variable should never be done. Questions 12 and 13 address this. Even though the poverty rate for Puerto Rico is fabricated, students can still practice using a regression equation to predict the response variable. If the poverty rate was in fact 25% and this was a valid value to use in the regression equation, the students would see that, $20.323 + .510(25) = 33.073$. So the obesity percentage in Puerto Rico would be 33.073 percent. In terms of the actual use of this prediction, students should adamantly refuse to use it in a real-life scenario. There are no data points providing information to the regression equation near a poverty rate of 25%. There is no evidence to suggest that the linear relationship will continue past the last poverty rate in the data set.

IV. Interpret the Results

Throughout the activity, students have been asked to provide interpretations of results such as interpreting the parameter estimates of the model and the correlation coefficient or explaining outliers, residuals, and the model SSE. One important remaining question is left for question 14, which asks students if the regression equation implies that poverty rate directly causes the increased obesity percentage in states across the United States. Clearly, students should understand that correlation between the two variables in no way implies causation. This study was purely observational, so the students were simply asked to determine if there was a relationship between poverty and obesity. This does not check to see if poverty leads to obesity because an experimental study would have needed to be created that took subjects of similar characteristics, divided them into two groups with one living a period of time in severe poverty and the other living a comfortable lifestyle, and over the course of the study determined that the incidence of obesity in the poverty group is significantly higher compared to the group not living in poverty. Not only is this study unethical, but also nearly impossible to carry out. The only main difference between the two subject groups would have to be the income/lifestyle characteristic. Many confounding variables will play a role in the study. Therefore, students should understand that even though poverty and obesity are related, poverty does not cause obesity.

Assessment

1. Each month, the owner of Cafe Gardens restaurant records y = monthly total sales receipts and x = amount spent that month on advertising, both in thousands of dollars. For the first four months of operation, the observations are as shown in the following table.

Advertising	Sales
0	4
1	7
2	8
5	9

Calculate the following:

(a) Estimated regression equation.

(b) Correlation Coefficient.

(c) Residuals and SSE.

(d) Interpret the estimated slope term and the correlation coefficient.

2. Suppose an experiment involving 5 subjects is conducted to determine the relationship between the percentage of a certain drug in the bloodstream and the length of time it takes to react to a stimulus. The results are shown in the table below.

Amount of Drug x (%)	Reaction Time y (sec)
1	1
2	1
3	2
4	2
5	4

(a) Create a scatterplot for this data.

(b) Determine the regression equation using a calculator.

(c) Suppose a patient had 2.5% of the drug, what is their predicted reaction time in seconds?

Answers

1. (a) estimated regression equation: $\hat{y}_{sales} = 5.286 + .857(advertising)$

(b) $r = .8571$

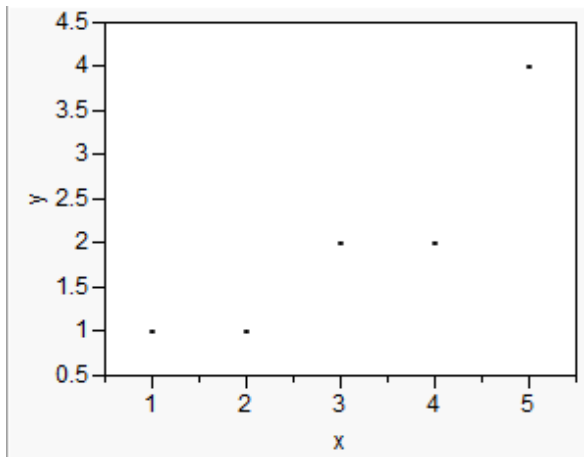
(c)

Advertising	Sales	residual
0	4	-1.286
1	7	.857
2	8	1
5	9	-.571

$$SSE = 3.714$$

(d) For every increase in \$1,000 spent monthly on advertising, we can expect the monthly sales to increase by \$857. The correlation coefficient of .8571 suggests a quite strong, positive, linear relationship between monthly advertising spending and monthly sales.

2. (a) A scatter plot for the data should resemble



(b) The estimated regression equation is: $\hat{y}_{time} = -.1 + .7(drug)$

(c) The predicted reaction time for a subject with 2.5% of the drug is $-.1 + .7(2.5) = 1.65$ seconds.

Potential Extension

In the activity, students found that Colorado and the District of Columbia might be outliers that are influencing the regression equation. Have the students explore how the regression equation changes if either both of the points are removed, or just one of them is removed. This will illustrate the influence of outliers.

References

1. Background on thematic maps adapted from the following sources:
<http://chnm.gmu.edu/worldhistorysources/unpacking/mapsmain.html>
http://en.wikipedia.org/wiki/Thematic_map
http://en.wikipedia.org/wiki/Geographic_information_system
http://en.wikipedia.org/wiki/Map_making
<http://www.phil.uni-passau.de/histhw/tutcarto/english/index-frames-en.html>
<http://www.ncjrs.gov/html/nij/mapping/>
2. Data Source: <http://www.irp.wisc.edu/faqs/faq3/table1.htm>
3. United States of Obesity Map Source: <http://calorielab.com/news/2011/06/30/fattest-states-2011/>
4. Assessment questions from: *Mind on Statistics*. Fourth Edition by Utts/Heckard, 2012. Cengage Learning.

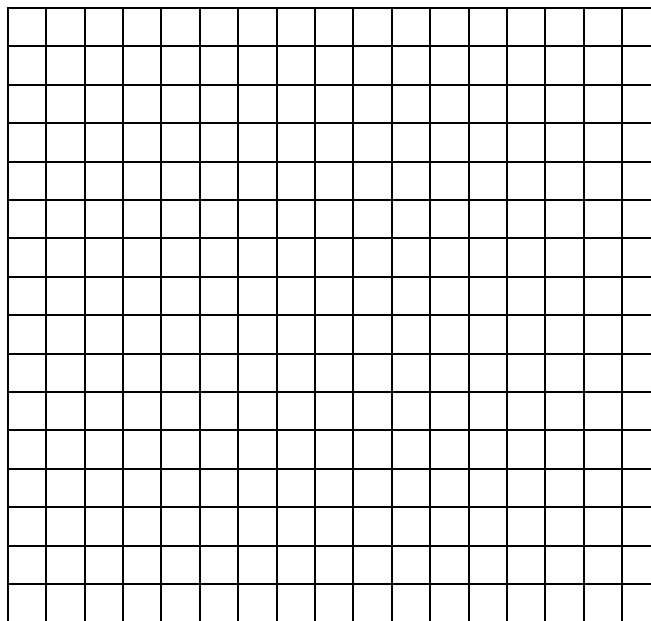
The United States of Obesity Activity Sheet

Activity Background

A thematic map is a map of a theme or topic. There are numerous varieties of thematic maps. Most of the maps utilized by media outlets such as a map that shows the status of sales taxes by state or a map that shows the world population density are thematic maps. A thematic map displays the spatial pattern of a theme or series of attributes. Thematic maps emphasize spatial variation of one or a small number of geographic distributions. These distributions may be physical phenomena such as climate or human characteristics such as population density and health issues. Location, of course, is important to provide a reference base of where selected phenomena are occurring.

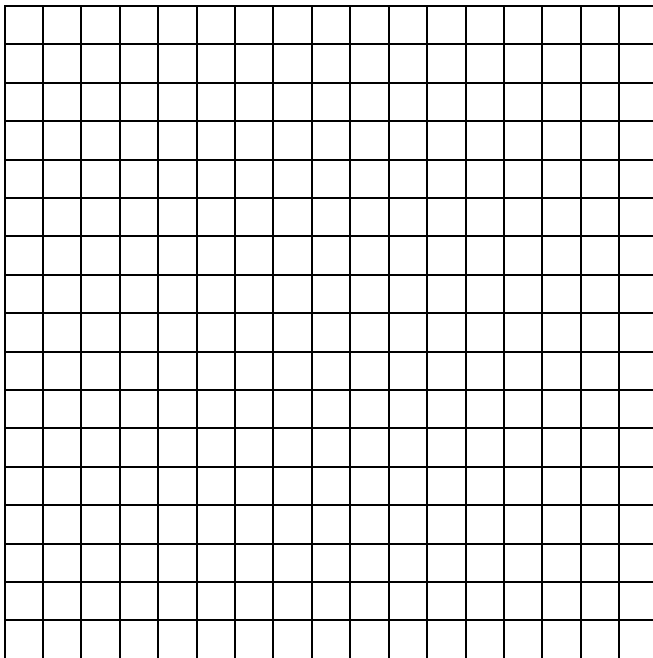
A choropleth map is a special case of a thematic map. Choropleth maps are particularly suited for charting phenomena that are evenly distributed within each set area. Raw data, for example a population distribution, should not be mapped with a choropleth map. However; if a derived value can be obtained from raw data (such as population densities), then the choropleth map can apply. To produce a choropleth map, the observations are grouped into a set of classes based on their data values and then each class is shaded with an appropriate color. Choropleth maps are one of the simplest methods of representing data that have been collected for areal units. They are commonly used to map census demographic data based on townships or census divisions. In this activity, you will apply topics from simple linear regression to explore the relationship between poverty and obesity in the United States between 2008 and 2010. Using the data provided at the end of this worksheet, complete the following questions.

1. Construct a scatterplot with each state's obesity percentage on the vertical axis and each state's corresponding poverty rate on the horizontal axis.



7. Based on the regression equation and scatter plot, are there any states that appear to be 'pulling' the line in one direction or the other? If so, which states are they, and in what direction to they appear to be pulling the line?

8. Using the estimated regression equation, calculate the residuals and then plot a histogram AND stem plot of the residuals.



9. Based on the histogram and stem plot of residuals in question 8, are there any states that may be outliers? If so, what states are they? What can you conclude about these states from your answer to question 7?

10. Using the residuals, determine the Sum of Squares for Error for the estimated regression equation. Note: $SSE = \sum residual^2$

11. Suppose you were to fit another regression equation for this data, but this time just by guessing the best-fit line. Would you expect the new SSE to be larger, smaller, or equal to the value found in number 10? Justify your answer.

12. Suppose Puerto Rico became a state and it had a poverty rate of 25%. Use the regression equation found in question 4 to predict the obesity percentage for Puerto Rico.

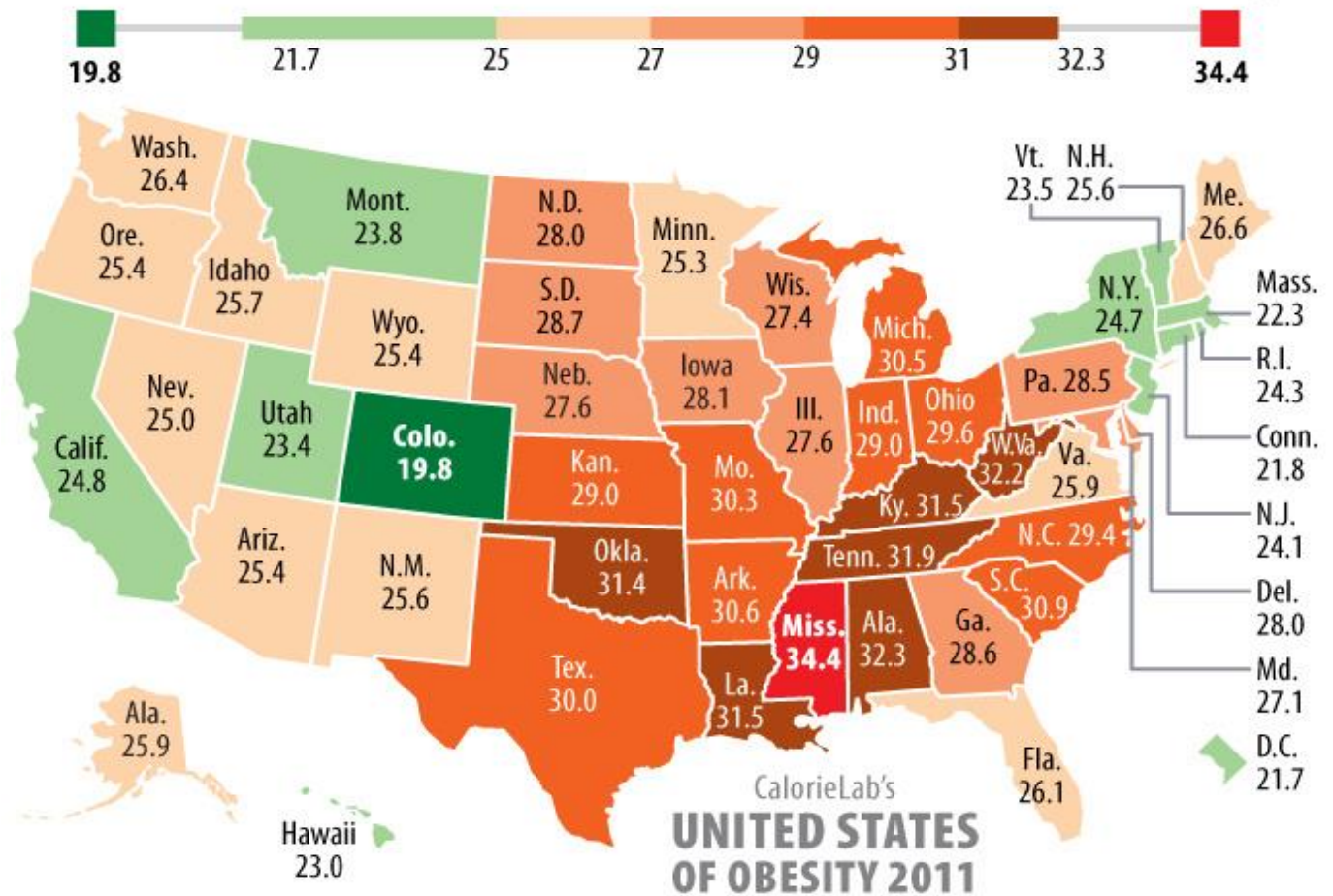
13. Is it okay to extrapolate the regression equation to include Puerto Rico? That is, is it reasonable to apply the regression equation to predict the obesity percentage for Puerto Rico?

14. From this simple linear regression analysis, can you conclude that an increase in the poverty rate directly causes the obesity percentage to increase? Why or why not?

Leanest State
Colorado

Percentage of Obese Adult Population
(3-year average from 2008-10 CDC Behavioral Risk Factor Surveillance System data)

Fattest State
Mississippi



State	Poverty Rate	Obesity Percent	State	Poverty Rate	Obesity Percent
Alabama	16.1	32.3	Missouri	14.6	30.3
Alaska	10.8	25.9	Montana	13.4	23.8
Arizona	19.2	25.4	Nebraska	10.2	27.6
Arkansas	16.5	30.6	Nevada	13.4	25
California	15.4	24.8	New Hampshire	7.1	25.6
Colorado	11.9	19.8	New Jersey	9.8	24.1
Connecticut	8.3	21.8	New Mexico	19.1	25.6
Delaware	11.3	28	New York	15.3	24.7
District of Columbia	18.1	21.7	North Carolina	16.1	29.4
Florida	14.6	26.1	North Dakota	11.6	28
Georgia	17.5	28.6	Ohio	14.1	29.6
Hawaii	11.5	23	Oklahoma	14.3	31.4
Idaho	13.3	25.7	Oregon	12.7	25.4
Illinois	13.2	27.6	Pennsylvania	11.4	28.5
Indiana	15.6	29	Rhode Island	13.1	24.3
Iowa	10.2	28.1	South Carolina	14.9	30.9
Kansas	13.6	29	South Dakota	13.5	28.7
Kentucky	17.3	31.5	Tennessee	16.1	31.9
Louisiana	18	31.5	Texas	17.2	30
Maine	12	26.6	Utah	9.1	23.4
Maryland	9.7	27.1	Vermont	9.7	23.5
Massachusetts	10.9	22.3	Virginia	10.6	25.9
Michigan	14.2	30.5	Washington	11.2	26.4
Minnesota	10.5	25.3	West Virginia	15.7	32.2
Mississippi	21.3	34.4	Wisconsin	10.2	27.4
			Wyoming	9.6	25.4