Mathematical Modeling and Consulting

Sponsor

The RAND Corporation

Progress Report

Graph-Theoretic Variables to Explore the Effects of Globalization on Obesity

Team Members

Michael Weinberger, michael.lee.weinberger@gmail.com

Zhendan Zhu, zhendanzhu@hotmail.com

Shannon Cebron, scebron@cis.jhu.edu

Academic Mentor

Dr. N. .H. Lee, Applied Mathematics and Statistics

nhlee@jhu.edu

Date: Last Compiled on December 19, 2012

Disclaimer: This is a class project whose findings are endorsed by neither Johns Hopkins University nor the RAND Corporation.

Abstract

In recent years, obesity has become a global epidemic, affecting not only citizens of first-world countries, but also citizens of those countries experiencing rapid trade development. The link between globalization and obesity has been explored by some organizations, but not in a rigorous mathematical context. This project seeks to develop a mathematical description of this uptake in obesity, utilizing graphical constructs based in known information about countries development rates and trading habits, and using properties of these graphs to build regression models which can accurately predict obesity rates.

Contents

List of Figures

Chapter 1

Introduction

Obesity is a medical condition identified by a Body Mass Index (BMI) (an adjusted proportion between height and weight) greater than 30. Obesity is proven to have extreme effects on a person‘s quality of life [3]. It is a major predictor in several potentially deadly types of disease; a common cause of physical degradation in the body manifested through pain in the joints and difficulty walking and moving; an aggravator of existing conditions such as sleep apnea and acid reflux disease; and a known detriment to mental health for reasons of body image and self-confidence.

When present in large proportions in populations, obesity is a major public health problem and a leading cause of preventable death. Treatment for heart disease, asthma, and diabetes is costly, driving up the cost of health care for the whole population. Furthermore, lost work hours due to obesity-related health problems detract from the health of the local economy.

As of 2010, the United States Center for Disease Control reports that 35.7 percent of the American population is obese, and this number has been steadily increasing since the 1960s [1]. The federal government estimates that up to USD 117 billion is lost yearly due to direct and indirect costs of such an obese population [1].

The prevalence of obesity is increasing in countries around the world, at the highest rates in countries recently achieving highly developed status as according to the Human Development Index (a weighted average exceeding 0.8 between measures of education, life expectancy, and personal wealth). However, in some countries this effect is more pronounced than in others, and at this level of detail, the spread of obesity is not well-understood. It is surmised that through increased trade and increased personal disposable income, more processed food has become available around the world. Furthermore, rapid changes in technology have allowed nations to become more culturally integrated with one another, and some experts suggest that citizens of developing countries are becoming more and more influenced by the dietary and exercise habits of their developed neighbors.

A better understanding of the interplay between development, globalization, and obesity may contribute positively to efforts to prevent the spread of this preventable, expensive, and deadly disease.

The RAND Corporation is a public policy thinktank which conducts research and analysis to support political decision-making and the public good. Many elected persons agree that legislation may help to curb the obesity epidemic, and it is important that this legislation is designed in an informed, rigorous manner. By working with the RAND Corporation to conduct this project, we can help to give lawmakers insight into what factors most prominently affect obesity trends.

Chapter 2

Technical Background

Regression Analysis

Regression analysis is a statistical technique for estimating the relationships between different variables. A standard regression model consists of a single response, or dependent variable; and one or more predictor, or independent variables. The purpose of such a model is to predict how changes in these independent variables elicit changes in the response variable.

More specifically, a regression model relates the response variable, Y, to a function of the vector of independent variables, X, and their corresponding coefficients vector, β. The prediction function has the form $\widehat{Y} = f(X, \hat{\beta}) + \varepsilon$.

Several underlying assumptions come with performing regression analysis. First, we must assume that the sample we are using is representative of the population for which the inference will be useful. Since our sample data consists of nearly every country, our model meets this assumption. Second, we must assume that the errors for each prediction instance are independent from one another and from the value of the independent variables - more specifically, $E[\varepsilon_i|x_i] = E[\varepsilon_i] = 0$. We can control for meeting this assumption by examining data on the residuals of our model - the differences between the predicted values and actual values for each instance. If the assumption is met, then these residuals will follow no particular pattern, appearing to be random and uncorrelated with any other factors.

Third, we assume that the predictors are not correlated with one another; that is, no two independent variables are telling us the same information and thus clouding the quality of our model. We can detect this by investigating a covariance matrix among all independent variables, and taking action to modify, combine, or remove such variables that have a high correlation with another variable. The current standard is that a correlation value between two independent variables of 0.4 or greater warrants further investigation. Variables with a correlation value of 0.8 or greater are almost always a problem. The consequences of this situation are that the affected coefficients have larger standard errors than they otherwise would - our estimates are less precise; the model tends to overfit the data; and some computer programs may produce numerically inaccurate estimates because the matrix $X^T X$ is either not invertible, or not invertible at the level of accuracy that the software has. Some standard remedies for the problem of multicollinearity include dropping one of the variables, mean-centering the variables, or creating a new variable which represents both variables together. In our project we used the first and third of these remedies to deal with two different pairs of variables.

Finally and most importantly, for a linear regression model, it is assumed that each

predictor variable has a linear relationship with the response variable, which in our case, is obesity. Some variables may appear to have logarithmic, exponential, polynomial, or even more complicated relationships with the response variable. There are standard variable transformations to deal with these more complicated relationships and manifest them as linear ones. For example, if it appears that the obesity rate varies closely with the square root of a variable, then we transform that variable into its square root for inclusion in our regression analysis. This new variable can be equally informative to the original version, while also meeting the assumptions demanded by the regression model.

For multiple linear regression, $\hat{\beta}$, the vector of estimated slope coefficients, is obtained through maximum likelihood estimation, which serves to find the set of coefficients $\hat{\beta}$ that minimizes the sum of squared differences between the actual and predicted values. Minimizing squared differences rather than absolute differences allows us to penalize progressively more those observations for which the prediction error is very large. More specifically, in utilizing maximum likelihood estimation, we give a function which gives the probability that the obesity rates are equal to their current values, given that the set of coefficients is $\hat{\beta}$. The goal is to maximize this likelihood, which implies the greatest possible accuracy. We do so by taking the derivative of this function and solving for $\hat{\beta}$ when the function is equal to 0.

In the opposite of the case when two variables give us redundant information, it is also possible that two variables jointly give us more information then they do individually. In this case, we include an interaction term in our regression model, where the variables multiplied together is considered an additional independent variable. A coefficient is also estimated for this interaction term, just as for any other term in the model. With so many variables present in our obesity model, it is easiest to use software to step through all of the possible interactions and find out which ones contribute significantly to the strength of the model.

In terms of the model's strength, there are a number of criteria which it must meet. We conduct an F-test on the regression model, which tests the hypothesis that the model fits the data well. A p-value of less than 0.05 will be sufficient in this case. We also conduct a t-test on each individual coefficient estimate, which tests the hypothesis that the coefficient is nonzero. If the t-test is significant, i.e., has a p-value of less than 0.05, then we have shown within reasonable doubt that there is a significant linear relationship between the predictor and the response. We are also concerned with the model's coefficient of determination, or R^2 value. The R^2 value gives an estimate of what percentage of the data's variability is accounted for by the regression model. The value can fall between 0 and 1, and we would seek out a value of at least 0.8.

Predictor variables which do not have a strong linear relationship with the response variable may add noise to the model and detract from its accuracy. In general, we desire a model that is as simple as possible. We can conduct ANOVA tests between nested models to see if the model with added terms (whether interaction terms or additional independent variables) is significantly stronger than the smaller model. If it is not, then we elect to use the simpler model that explains the data with the same efficacy. If the ANOVA test result has a p-value of less than 0.05 then we accept the larger model.

The final model that we will settle on will have low enough p-values in all of these criteria, and will by virtue of the ANOVA test be stronger than any of the simpler models that are possible. It will also meet the assumptions listed above which are inherent to the use of linear regression to describe relationships among data.

Vertex-edge Graphs and their Properties

In mathematics, vertex-edge graphs are data structures consisting of a collection of nodes (vertices), and arcs (edges) which connect pairs of nodes. Graph theory is of specific interest to our project because it allows for an abstraction of the relationships between world nations into a mathematical form that can be analyzed rigorously. Our primary goal in incorporating vertex-edge graphs into our project is to gather data from these graphs that can be used as potential independent variables in our regression analysis. The independent variables we use must exist as attributes of a country, and thus in terms of graph theory, we will specifically be interested in those graph features which characterize specific vertices, as this can be transformed into an independent variable corresponding to a country represented by the vertex.

The vertex characterizations which are of particular interest to us are as follows (please see the glossary for further definitions of these characterizations):

- Vertex degree
  - The incoming degree is the number of directed edges ending at the vertex
  - The outgoing degree is the number of directed edges beginning at the vertex
  - The overall degree is the total number of edges incident to the vertex
- Shortest path to a specific vertex; e.g., for each country, we calculate the length of a shortest path from its corresponding vertex to the vertex representing the United States
  - A weighted shortest path finds the shortest sum of edge weights leading to the target vertex
  - An unweighted shortest path finds the shortest number of edges on a path to the target vertex

All of those listed vertex attributes can be converted into continuous or factor variables for use as independent variables in our regression analysis. For example, the length of the weighted shortest path from each vertex to the vertex corresponding to the United States, where adjacencies are defined by border relationships and edge lengths are defined by the border's length in kilometers, was incorporated as an independent variable in our regression model. A simple linear regression model between this path length and the obesity rate demonstrated that obesity rate decreased as the length of the path increased. Thus we can infer that countries which are geographically close to the United States, and especially those with shorter border lengths, tend to have higher obesity rates.

Data and Software

Vast amounts of data for nations and sovereign entities are available from the CIA World Factbook and the World Health Organization's (WHO) databases. The data can be downloaded from these sources in .csv files. THese .csv files can be imported into the R environment and merged so that a data frame is built with several different independent variables listed for each country. The CIA World Factbook and WHO refer to some countries by different names (e.g., the Republic of Korea and South Korea are the same country, but

named differently by these sources). In order to merge the lists properly despite these naming conflicts, another table was generated which listed possible alternate names for each country, and referenced during the merging process to ensure that data for South Korea and the Republic of Korea, for example, was combined. As a convention, the name used by the CIA World Factbook was our default name.

The R environment is effective for statistics and data management, but does not have a robust package for creating the type of graphs discussed in the previous subsection. Python, however, does have such a feature. Therefore, the data from R must be converted into a .csv file again, which is easily done with a data frame by using the "write.csv" command in R. However, our data frame with different independent variables does not define country contingencies in any way - i.e., the data needed to define edges in our graphs. This information must be mined from a webpage and processed to remove commas, footnotes, and other extraneous text that the CIA has included in their notes. This was done in Microsoft Excel by pasting the contents of the webpage onto a spreadsheet and processing the cells using Excel formulas. This was then exported as a second .csv file and imported into Python along with the country names in order to create the graphs.

Python maintains a package called Pygraph which allows for relative ease with creation of graphs, and ease of data extraction, such as shortest path data (see the previous subsection). These data were easily arranged by country and exported back into R, for merging into our main data frame. After this step, we several dozen potential independent variables to consider in our regression analysis, when combining our initial 16 independent variables with another 20 independent variables that were extracted from our theoretical graphs. At this point we were ready to begin testing relationships and building regression models in R, based on the criteria explained above.

R also offers the opportunity for any user to create a package, which is a collection of R functions, data, and compiled code in a clearly-documented format. It is easy to include not only a developed regression model, accompanying data, and example code for diagnosing the model's strength or changing its parameters, but also to include images of our graphs, a glossary explaining independent variables in the model, a list of references, and more.

Chapter 3

Problem Statement

Obesity affects persons all over the globe, and as a result affects the ability of governments to most effectively provide services for their citizens. The phenomenon is growing, rather than subsiding. Without the development of effective tools for managing the obesity epidemic, we will see it continue to expand, much to the detriment of global health and sufficiency.

Large amounts of data exist on the prevalence of obesity. While there are many existing statistical tools to analyze these data, many spurs in obesity rate remain unexplained. The world continues to become more interconnected, with countries influencing one another like never before through their various relationships. An attempt to introduce a quantification of this influence effect is necessary in order to give a more robust interpretation of the growing obesity problem and allow policymakers to have more tools for combating it. Our sponsor, the RAND Corporation; the WHO; and other entities need more complex data modeling tools than ever.

Our goal is to make use of already-existing statistical tools to explore the obesity phenomenon, but through a new lens. We intend to incorporate complex data about the relationships between different countries in order to create a more statistically robust model, which as demonstrated by the results of an ANOVA test, is stronger and more informative than a model which does not include graph-theoretic variables. Therefore, this project will not only make use of results from the statistical community, but of the properties and algorithms of vertex-edge graphs. We must develop theoretical graphs using informed representations that will shed light on the complex problem of obesity. Our problem is not only to diagnose and remedy our model such that it is statistically sound, but to diligently develop our independent variables in the first place. Depending on how we define adjacencies and edge weights in the graphs that we examine, we may find more or less informative results. Thus it is paramount to adhere to the rules of the mathematical community while also being creative and adaptable to the results of our choices in constructing the graphs.

The process of this project consists of several major parts: acquiring data (both non-graphical independent variables and data on relationships between countries); constructing graphs and extracting information from them to add as independent variables; and analyzing these data and building an effective regression model. This regression model will have an F-statistic with a p-value of less than 0.05; p-values of less than 0.05 for the t-tests for each individual independent variable; an R^2 value of at least 0.8 for its efficacy in explaining the variation of the data; and will meet all linear regression assumptions. Furthermore, through the process of stepwise variable selection, we intend to find the best possible model meeting these criteria, by eliminating unnecessary variables.

As a result, we will be able to provide a regression equation which effectively predicts

a country's obesity rate based on its corresponding values of the independent variables. We will not only state this model, but will also provide an R package which contains all of our data, a function which computes the regression predictions based on independent variable values, as well as supplementary images and descriptions of included variables to assist in understanding of the model. Thus this R package will allow a user to estimate how a country's obesity rate would change based on changes in any number of the predictors. Therefore, our model may be an effective tool for policymakers wishing to gain a sense of the effect on the national obesity rate of certain policy decisions which would elicit changes in these variables.

Beyond being a useful tool for policymakers the model will also more simply make a contribution to an understanding of how obesity has spread throughout the world. The relationships demonstrated by the model may be of interest to global health analysts and historians, who will now potentially have mathematical ideas to back up their speculations as to the effects of globalization on obesity.

Chapter 4

Analysis

Non-graph-theoretic independent variables

The following independent, non-graphical variables were investigated during our model building:

- Total electricity consumption
- Fertility rate
- Youth unemployment
- Total population
- Per person imports (USD)
- Per person GDP (USD)
- Percentage of children under age 5 who are underweight
- Population growth rate
- Birth rate
- Education expenditures as a percentage of GDP
- Health expenditures as a percentage of GDP
- Inflation rate
- Gini index for families
- Total internet users

Relationships individually between these independent variables and the response variable, obesity rate, were explored for inclusion in a regression model. For the following variables, we were unable to find a linear relationship, or find a suitable transformation to elicit a linear relationship:

- Electricity consumption
- Youth unemployment

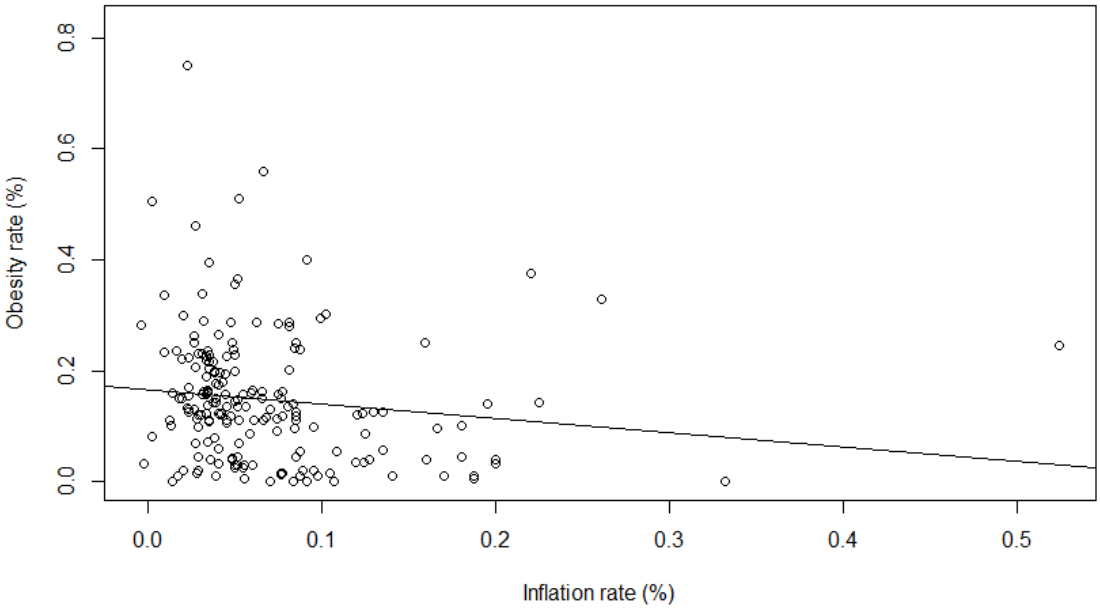

Figure 4.1: Example scatterplot of a lack of linear relationship, between inflation rate and obesity rate (the overlaid line is the simple regression line)

- Education expenditures as a percentage of GDP
- Health expenditures as a percentage of GDP
- Inflation rate
- Gini index
- Total internet users

These variables were removed from consideration in regression analysis.

The following are those variables which had a significant linear relationship, without any linearizing transformations required. They are listed along with the p-value from the t-test used to test the strength of the relationship.

- Fertility rate ($p = 6.35 \cdot 10^{-07}$)
- Total population ($p = 0.0521$)
- Per person GDP (USD) ($p = 0.00525$)
- Percentage of children under age 5 who are underweight ($p = 5.05 \cdot 10^{-15}$)
- Population growth rate ($p = 1.60 \cdot 10^{-06}$)
- Birth rate ($p = 1.43 \cdot 10^{-06}$)

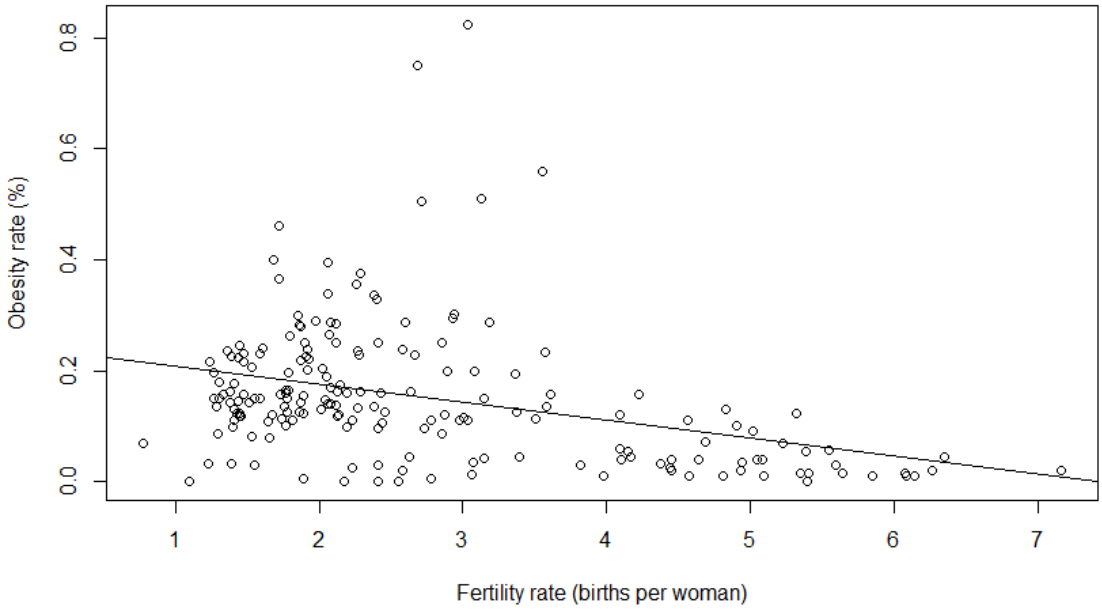

Figure 4.2: Example scatterplot of a strong linear relationship, between fertility rate and obesity rate (the overlaid line is the simple regression line)

Finally, some variables showed a stronger linear relationship following a transformation of some kind. This group of variables is also listed along with their transformations and their t-test p-values.

- Per person imports (USD)
  - Transformation: $f(x) = \log x$
  - $p = 2.70 \cdot 10^{-07}$
- Per person GDP (USD)
  - Transformation: $f(x) = \log x$
  - $p = 1.85 \cdot 10^{-07}$
- Percentage of children under age 5 who are underweight
  - Transformation: $f(x) = \sqrt{x}$
  - $p \leq 2.00 \cdot 10^{-16}$

Choice of graphs and graph-theoretic variables

We use three different graphs in order to create three theoretical graphs in Python.

Graph 1 was defined using a vertex corresponding to each country, and an undirected edge between countries in the event that the countries border one another.

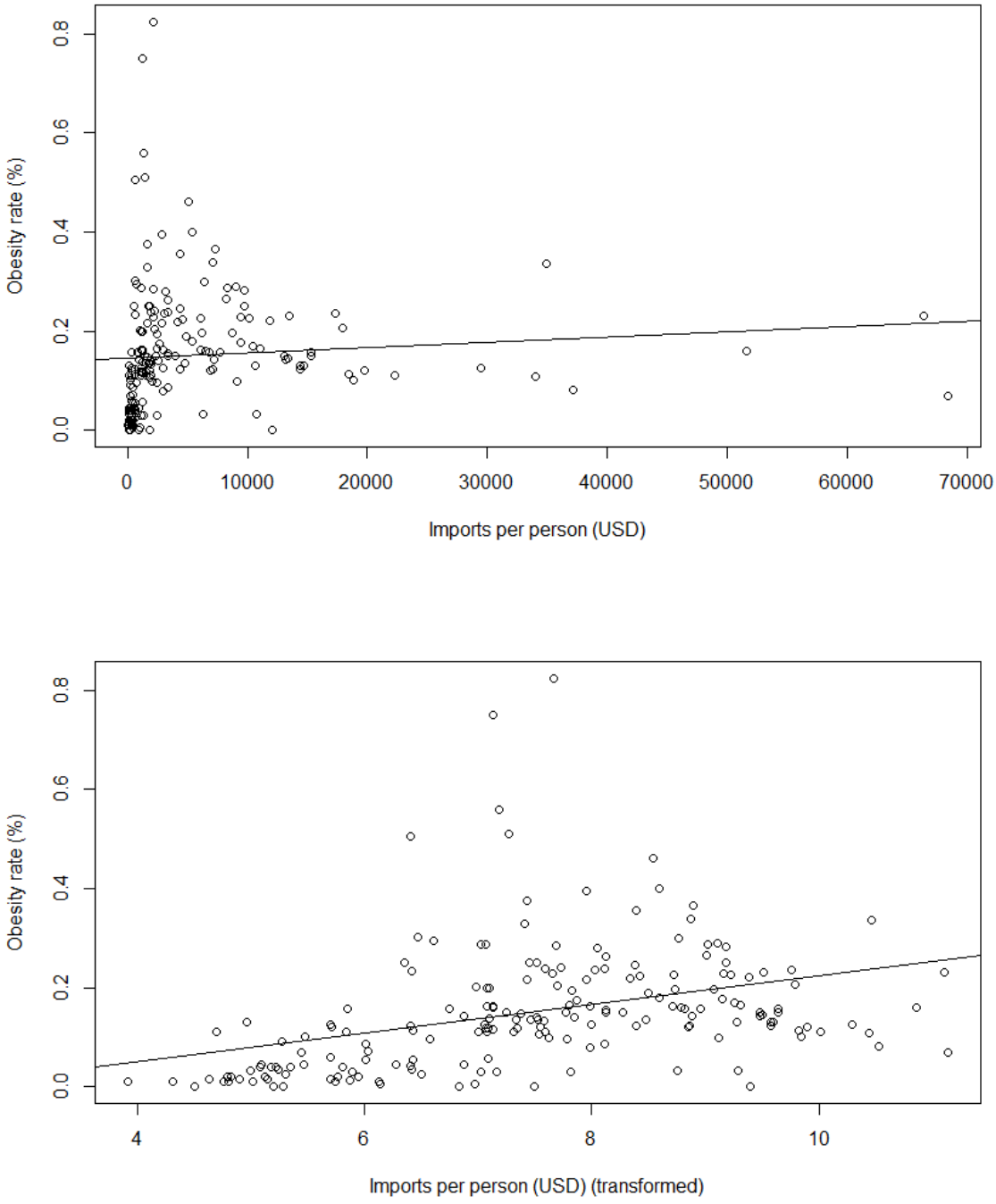

Figure 4.3: Plots of per person imports versus obesity rate, before and after the linearizing transformation (regression lines overlaid)

- *Weighting 1* was defined by giving existing edges a weight equal to the length of the border, in kilometers, between the two countries represented by their endpoints.
- *Weighting 2* was defined by giving existing edges a weight equal to the absolute difference between the obesity rates of the two countries represented by their endpoints.

Graph 2 was defined using a vertex corresponding to each country, and a directed edge from vertex i to vertex j if the country corresponding to vertix i imports over 4% of its total imports from the country corresponding to vertex j (i.e., lists the country as a major import partner).

- *Weighting 1* was defined by giving existing edges a weight equal to the specific percentage of imports.
- *Weighting 2* was defined by giving existing edges a weight equal to the absolute difference between the obesity rates of the two countries represented by their endpoints.

Graph 3 was defined using a vertex corresponding to each country, and a directed edge from vertex i to vertex j if the country corresponding to vertix i exports over 4% of its total exports to the country corresponding to vertex j (i.e., lists the country as a major export partner).

- *Weighting 1* was defined by giving existing edges a weight equal to the specific percentage of exports.
- *Weighting 2* was defined by giving existing edges a weight equal to the absolute difference between the obesity rates of the two countries represented by their endpoints.

The following variables were extracted for each vertex from our graphs in Python:

- *Graph 1*:
  - Total vertex degree
  - Shortest path to the United States, unweighted graph
  - Shortest path to the United States in terms of weighting 1
  - Average weight of incident edges in terms of weighting 2
  - Average obesity rate of all countries whose corresponding vertices are adjacent to the vertex in question
- *Graph 2*:
  - Number of incoming edges (i.e., the number of countries which have that country as a major source of imports) minus number of outgoing edges (i.e., the number of major import sources the country has)
  - Shortest path to the United States in terms of sum of weighting 1
  - Average weight of outgoing edges in terms of weighting 2
- *Graph 3*:
  - Number of incoming edges (i.e., the number of countries which have that country as a major target of exports) minus number of outgoing edges (i.e., the number of major export targets the country has)

  - Shortest path to the United States in terms of sum of weighting 1
  - Average weight of outgoing edges in terms of weighting 2

Relationships individually between these graph-extracted independent variables and the response variable, obesity rate, were also explored for inclusion in a regression model. For the following variables, we were unable to find a linear relationship, or find a suitable transformation to elicit a linear relationship:

- *Graph 3*: Shortest path to the United States in terms of sum of weighting 1
- *Graph 2*: Number of incoming edges minus number of outgoing edges
- *Graph 3*: Number of incoming edges minus number of outgoing edges

These variables were removed from consideration in regression analysis.

The following are those variables which had a significant linear relationship, without any linearizing transformations required. They are listed along with the p-value from the t-test used to test the strength of the relationship.

- *Graph 1*: Shortest path to the United States with unweighted edges ($p = 1.48 \cdot 10^{-05}$)
- *Graph 1*: Shortest path to the United States in terms of weighting 1 ($p = 4.48 \cdot 10^{-08}$)
- *Graph 1*: Total vertex degree ($p = 0.000491$)
- *Graph 1*: Average weight of incident edges in terms of weighting 2 ($p = 7.10 \cdot 10^{-10}$)
- *Graph 1*: Average obesity rate of countries whose corresponding vertices are adjacent to the vertex in question ($p \leq 2.00 \cdot 10^{-16}$)
- *Graph 2*: Average weight of outgoing edges in terms of weighting 2 ($p = 2.22 \cdot 10^{-11}$)
- *Graph 3*: Average weight of outgoing edges in terms of weighting 2 ($p = 1.81 \cdot 10^{-12}$)

Finally, one variable showed a stronger linear relationship following a transformation.

- *Graph 2*: Shortest path to the United States in terms of sum of weighting 1
  - Transformation: $f(x) = \log \log x$
  - $p = 0.00729$

Relationships among independent variables

Multicollinearity

We calculated correlation values between all pairs of independent variables as a check for the effect of multicollinearity on our model. The correlation between fertility rate and birth rate was 0.9703, compared to a maximum value of 1. The relationship between the two is evident from a scatterplot.

As discussed in the previous chapter, this relationship, known as multicollinearity, means that the two variables are offering redundant information to our model. For this particular case we were able to find an easy solution. We created a new independent variable equal to

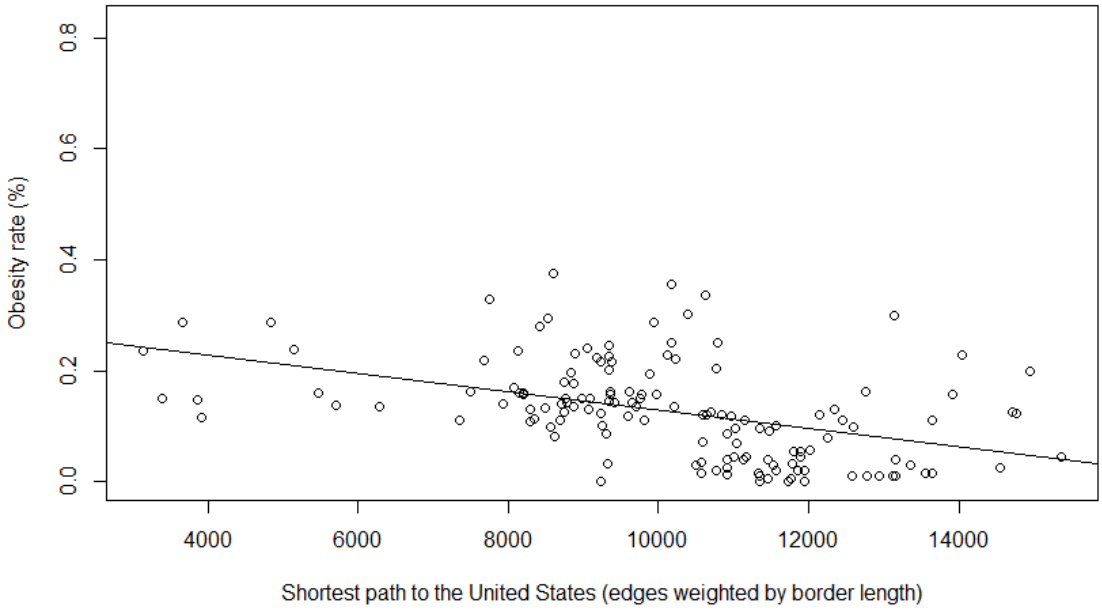

Figure 4.4: Example scatterplot of a strong linear relationship, between shortest path to the United States and obesity rate (the overlaid line is the simple regression line)

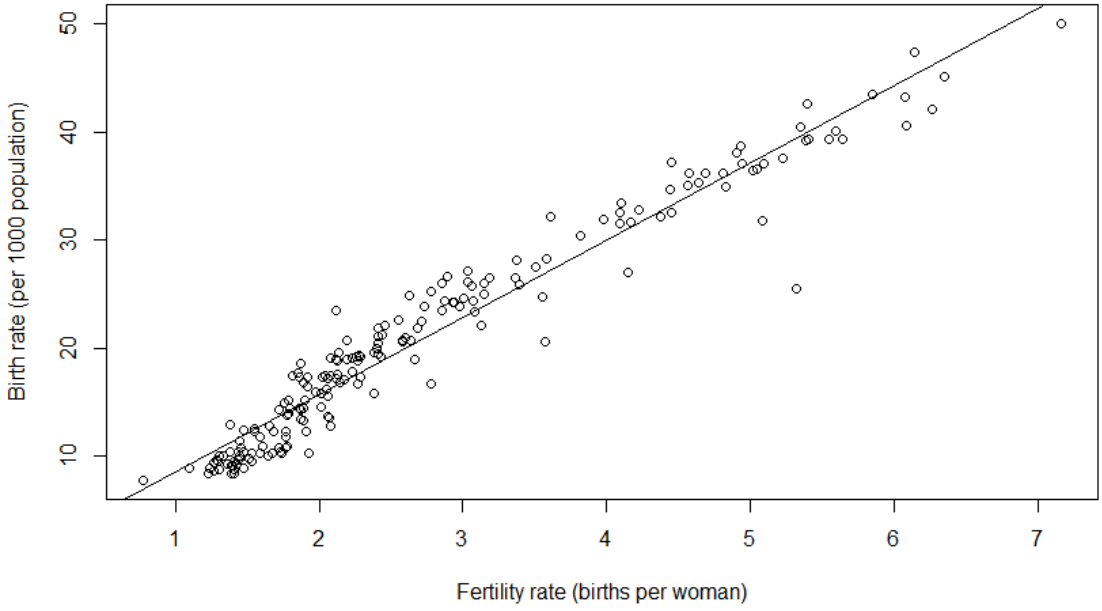

Figure 4.5: Scatterplot of fertility versus birth rate, overlaid with a simple regression line

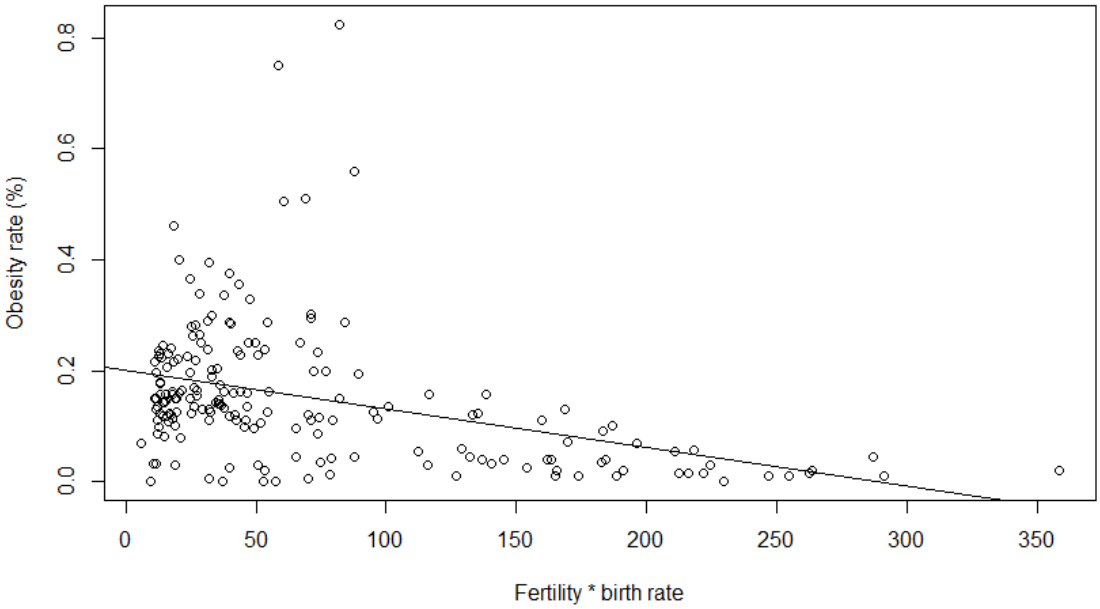

Figure 4.6: Scatterplot of the new variable versus obesity, overlaid with a simple regression line

the fertility rate multiplied by the birth rate. A hypothesis test on the linear relationship of this new variable with obesity rate gave $p = 3.29 \cdot 10^{-08}$, which is even smaller than the p-values for the variables individually.

There was also a very strong correlation between two variables extracted from Graph 1: shortest path to the United States in terms of number of edges, and in terms of sum of edge lengths when weighted by national border length. These two variables had a correlation of 0.9050. In this situation, we did not find that combining the two variables into a new variable was able to yield a stronger relationship than the variables on their own. Therefore, we dropped from our analysis the variable which had a weaker linear relationship with obesity rate, which was the shortest path in terms of the number of edges.

Missing values

In particular, one variable found to be statistically significant was missing data for many of the included countries. Many highly developed countries no longer collect or publish data on the percentage of underweight children in their country, because it tends to be very small. For this reason, we had to remove this variable from our model. It restricted the number of countries we could consider and thus reduced the strength of relationship of obesity rate with other predictors.

Stepwise regression to determine final variable inclusion

Some individual variables which show strong linear relationships in a simple regression against obesity rate, do not show strong statistical tests for their coefficients when included in a multiple linear regression model among all individually significant independent variables. Our stepwise regression was performed using elimination in both directions, comparing models based on their AIC. Out of the 12 independent variables considered in multiple linear regression, 5 were removed based on the results of stepwise regression.

Validation of usage of graph-theoretic variables

Finally, we used an ANOVA test between a multiple linear regression model created through stepwise regression *without* any of the graph-theoretic, and a multiple linear regression model created through stepwise regression *including* all significant graph-theoretic variables. This ANOVA test tests the hypothesis that the model is stronger when the added variables are included. Our test was statistically significant ($p \leq 2.00 \cdot 10^{-16}$), supporting our hypothesis that graph-theoretic variables offer added insight beyond non-graph-theoretic ones.

Chapter 5

Results

After analyzing all of the potential independent variables, checking for linear regression assumptions, and verifying the significance of all relevant hypothesis tests and statistics, we developed a final regression model:

$$\widehat{Y} = f(X, \hat{\beta}) + \epsilon$$

where

$$X = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ \log x_7 \end{pmatrix}$$

and $\beta = \begin{pmatrix} 1.11 \cdot 10^{-2} \\ -3.10 \cdot 10^{-4} \\ -1.60 \cdot 10^{-10} \\ -1.10 \cdot 10^{-5} \\ 5.23 \cdot 10^{-1} \\ 4.97 \cdot 10^{-1} \\ -3.04 \cdot 10^{-3} \\ 1.64 \cdot 10^{-2} \end{pmatrix}$

In the appendix, a final list of variables corresponding to the vector X and detailed explanations of their meanings can be found.

The F-test for our model's overall ability to fit the data gave a p-value of $p \leq 2.00 \cdot 10^{-16}$, suggesting that our model fits the data quite well. The R^2 value for our model was 0.7451, which suggests that our model explains most of the variation in the data.

Chapter 6

Conclusion

Our linear regression model is able to quite accurately predict the obesity rate of the 190 countries included in our analysis. We stated that a strong model should have an F-test p-value of less than 0.05. The value was $p \leq 2.00 \cdot 10^{-16}$, so this criterion was met. As discussed in chapter 4, each individual independent variable included in the model showed a significant linear relationship with the response variable. We specified that our goal was to achieve an R^2 of at least 0.8. Our value was 0.7451, so unfortunately, this goal was not met. However, our value was very close. This small deficiency should be a focus of future model development. The inability of our model to achieve a better R^2 value can possibly be attributed to a lack of other independent variables which explain some variation in a country's obesity rate. This motivates future exploration of graph-theoretic variables and other variables alike. Finally, thanks to the stepwise regression process, we have ensured that our model is the strongest possible model, and the simplest possible model achieving such strength. Our model also meets other assumptions of linear regression. A plot of residuals against fitted values shows that residuals have relatively no distribution in terms of the predicted values.

In Appendix C can be found comments on the individual relationships between predictor variables and obesity rate. In addition to developing this successful regression model, we have also produced an R package which has a function to output an obesity rate prediction based on specified values of the independent variables. The package also includes the data frame we used in our analysis, and several example plots and images. In this package can also be found our personal resumes and contact information, as well as descriptions of the independent variables used. Therefore, aside from the slightly smaller-than-desired R^2 value, we have met the demands set forth by our promised deliverables.

Future research should focus on two things.

1. Maintenance of updated information on predictor variables is necessary in order for the model to remain effective. Furthermore, completing missing data on percentage of underweight children in many countries would most likely strengthen the model significantly. A possible goal might be to find and record this information and use it to make the model more robust.

2. Gathering of obesity rate data for missing countries is also desirable. The world has more than 190 countries, but some countries do not collect obesity data or publish it. Inclusion of these data would impove the model's accuracy. A future researcher on this topic may also consider using overseas territories as separate instances in the model. For example, in this model, French Guiana was considered to be part of

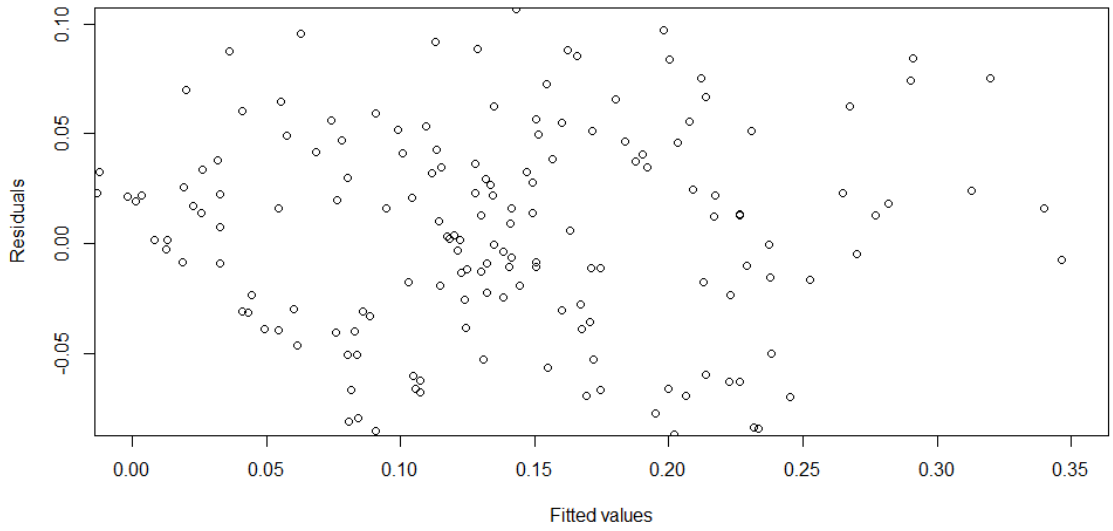

Figure 6.1: Scatterplot of residuals versus fitted values in our final regression model

France. However, French Guiana is in South America as opposed to Europe, and has a somewhat autonomous economy, political system, and culture. Therefore, it might be beneficial to use separate data for French Guiana.

Appendix A

Glossary

Obesity. A medical condition that is defined as having a Body Mass Index of greater than 30.

Body Mass Index. Body Mass Index is defined as the individual's body mass in kilograms, divided by the square of his or her height in meters.

Vertex-edge graph. A set of vertices (also called nodes), along with a set of edges (also called arcs), in which each edge must correspond to two vertices as its endpoints. The edge may or may not have direction. A simple graph is one that does not allow multiple edges between the same two vertices, or an edge that has the same vertex as both of its endpoints. We have only used simple graphs.

Human Development Index. A composite statistic of life expectancy, education, and income indices to rank countries into four tiers of human development.

Vertex degree. In graph theory, the degree (or valency) of a vertex of a graph is the number of edges incident to the vertex, with loops counted twice. The incoming degree of a vertex is the number of directed edges with the vertex as their ending point. The outgoing degree of a vertex is the number of directed edges with the vertex as their starting point.

Path. In graph theory, a path is a sequence of vertices, such that from each vertex in the sequence there is an edge to the next vertex in the sequence. No vertex may occur more than once in the path. In particular, a shortest path between vertex i and vertex j is the path such that the sum of the edge weights of edges included in the path is minimal. This path can be found using a number of algorithms, but most popular is Djikstra's algorithm. When edges in a graph are unweighted, for purposes of finding the shortest path each edge is considered to have a weight of 1; thus the length of the shortest path amounts to the number of edges that the path contains.

Regression. Regression is a method for modeling the relationship between a single response variable and one or more independent variables which are thought to have a relationship with the response variable. If this relationship is linear, then it is linear regression. If the response variable comes from a finite set of outcomes, then we use logistic regression, which calculates the odds of the response variable taking on each outcome. Coefficients are estimated for each independent variable such that the deviation between actual and predicted values is minimized.

Stepwise regression. Stepwise regression is a method of determining the inclusion of independent variables in a regression model through an automatic procedure, done in the form of a sequence of F-tests and information criterion comparisons, in order to find out which variables complicate the model without contributing significantly to its strength. Backward elimination involves starting with all candidate variables and testing the deletion of each successive variable, deleting any variables whose removals improve the model or do not detract from the model.

Residuals. In regression analysis, a regression model's residuals are the squared distances between actual values of the response variable and the the values predicted by the model. Generally, if the residual values appear to follow a pattern, rather than being randomly dispersed, we might surmise that the current model is not an appropriate choice for modeling the data.

Hypothesis test. A hypothesis test compares a null hypothesis to an alternative hypothesis. The null hypothesis is rejected if, under the assumption that it is true, the probability of observing the true data is less than a specified significance value. In the statistics community it is standard to use 0.05 as this value.

Gini index. A probabalistic measure of dispersion of incomes in a society, where a higher Gini index indicates more income inequality - i.e., a bigger gap between rich and poor.

.csv file. A file that stores tabular data in plaintext format, and can be easily processed in most programming languages. Rows are separated by line breaks, and row entries are separated by commas, tabs, semicolons, or other user-specified delimiters.

AIC. A measure of linear model performance which assesses the model's fit based on the amount of information lost when the model is used to describe reality. Absolute values are somewhat meaningless, but the AIC is effective for comparing multiple models.

Appendix B

Acronyms

CIA Central Intelligience Agency
GDP Gross Domestic Production
BMI Body Mass Index
WHO World Health Organization
CSV Comma-separated Values
USD United States Dollars
AIC Akaike information criterion

Appendix C

Regression variable descriptions

X_1: Birth rate multiplied by fertility rate. The fertility rate is measured as number of births per woman, and the birth rate is measured as number of births per 1000 people. These two metrics were highly correlated, and thus a new metric was created by multiplying the two values together. We found that a higher birth rate and fertility rate corresponded to a lower obesity rate.

X_2: Population. This is simply the total population of the country. Countries with larger populations tended to have lower obesity rates in our model.

X_3: Distance to the vertex corresponding to the United States, on graph 1, in terms of weighting 1. This measures the shortest path from the vertex corresponding to the given country, calculated using Djikstra's algorithm where edge weights are equal to the length of the border in kilometers. The higher this metric was for a country, the lower the country's obesity rate was. This means that countries geographically distant from the United States have lower obesity rates. Minimizing the path length based on border length means that paths tended to avoid pairs of countries sharing long borders with another. A long border between two countries may imply more cultural closeness.

X_4: Import differential. This metric corresponds to the average absolute difference in obesity rates between the given country and those countries from which it imports a large percentage of its total imports. Countries with a higher such differential have higher obesity rates. In other words, countries which tend to import from countries with more divergent obesity rates from their own, tend to be those countries which are the most obese.

X_5: Export differential. This metric corresponds to the average absolute difference in obesity rates between the given country and those countries to which it exports a large percentage of its total exports. Countries with a higher such differential have higher obesity rates. In other words, countries which tend to export to countries with more divergent obesity rates from their own, tend to be those countries which are most obese.

X_6: Number of borders. This was the vertex degree in graph 1 and is simply equal to the number of countries that the given country borders. Countries with fewer borders enjoy higher obesity rates. The strength of this relationship is probably influenced largely by the tendency of island nations to have high obesity rates. Other explanations exist, including the fact that landlocked countries inherently have more borders, but also tend to be economically disadvantaged due to not having access to any ports.

X_7: GDP per capita. This is measured in USD and is the country's total GDP divided by the country's population. When the log of this metric is taken, a higher value corresponds

to a higher obesity rate. This is intuitive; wealthier countries are more obese.

Selected Bibliography Including Cited Works

[1] Physical Activity CDC Division of Nutrition and Obesity. Adult obesity facts. http://www.cdc.gov/obesity/adult/causes/index.html Accessed Nov, 2012.

[2] CIA World Factbook. Obesity - adult prevalence rate. https://www.cia.gov/library/publications/the-world-factbook/fields/print_2228.html Accessed Nov, 2012.

[3] Stanford Hospital and Clinics. Health effects of obesity. http://stanfordhospital.org/clinicsmedServices/COE/surgicalServices/generalSurgery/bariatricsurgery/obesity/effects.htm Accessed Nov, 2012.