Mathematical Modeling and Consulting

Sponsor

The RAND Corporation

Progress Report

Large Graphical Models to Explore the Effects of Globalization and Development on Prevalence of 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 Complied on December 9, 2012

This project was supported by Johns Hopkins University.

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

Abstract **2**

1 Introduction **4**

2 Technical Background **5**

3 Problem Statement **9**

4 Analysis **10**

5 Results **16**

A Glossary **17**

B Acronyms **19**

REFERENCES

Selected Bibliography Including Cited Works **20**

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

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 control for this by investigating a covariance matrix among all independence 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 certainly causing multicollinearity in the model. 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.

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.

Alternatives to a linear regression model existed. More specifically, with the tools available, we had the opportunity to choose between linear regression, Gaussian regression, logit regression, Poisson regression, Gamma regression, inverse Gaussian regression, and quasipoisson regression. However, the other types of regression in this list assume a different structure of the data and a different presumed relationship between the independent variables. For example, Poisson regression deals with counting data, and expected counts based on certain factors. Estimating the count of obese people in a country is not an appropriate regression task because our data do not meet the assumptions of a Poisson model. Quasipoisson regression is a modified form of this. Gaussian and inverse Gaussian regression concern estimates of functions that must pass through specific sets of points. Logit regression estimates the odds of an outcome, essentially acting as a binary classification model. Gamma regression is a more generalized form of Poisson regression. None of these types of regression except linear are an appropriate fit for our data and goal.

For multiple linear regression, estimates for B are obtained through maximum likelihood estimation, which serves to find the set of coefficients B 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 B. 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 B 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 R2 value. The R2 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.

Theoretical Graphs

In mathematics, graph theory is the study of graphs, which are 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 theoretical 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):

- Vertex degree
- 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
- Membership in a cycle
- Membership in a maximal clique
- The condition of being a cut vertex
- The condition of being the endpoint of a cut edge

All of those listed vertex attributes can be converted into continuous or factor variables for use as independent variables in our regression analysis. The specific graphical independent variables used in our analysis will be discussed in later sections.

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

Consider the following graph.

This graph was constructed using major US cities as vertices and the distance in miles between the cities as edge lengths. Though the vertices in our model will represent nations, not cities, this is one example of how the graph might be structured.

The problem with inferring information from the graph is choosing how to define adjacencies, edge lengths, and vertex weights; in addition, it must be decided which graphical properties will then be examined in a statistical model. Our task is to use sequential regression modeling to investigate which properties are important predictors of national obesity rates, and use this information to inform our graph definitions and our final statistical models.

Chapter 4

Analysis

Non-graphical independent variables

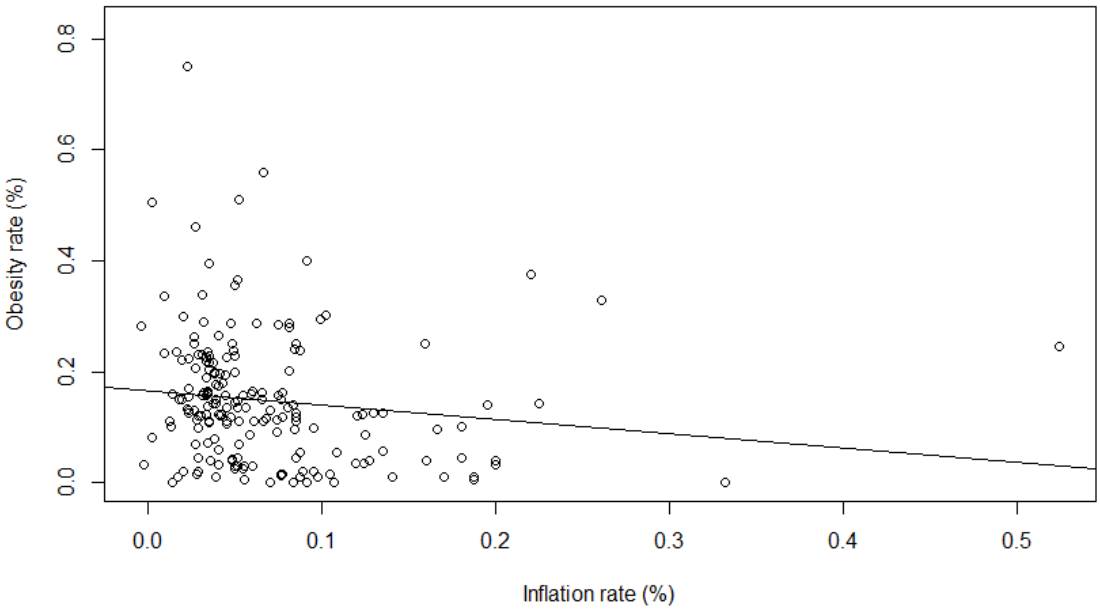
The following independent, non-graphical variables were investigated during our model building:

- 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)



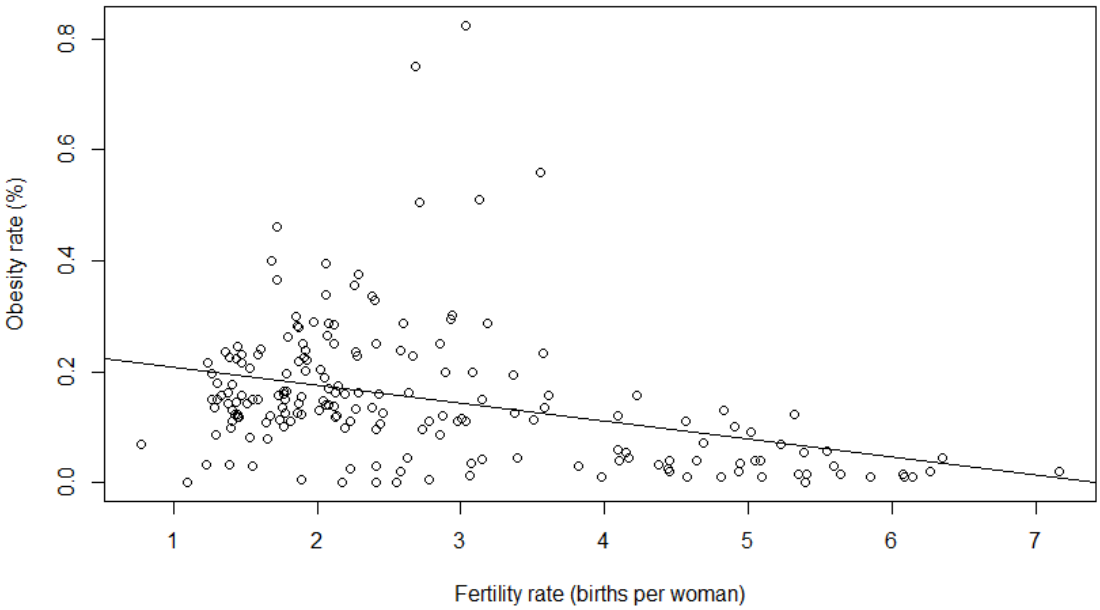
- Education expenditures as a percentage of GDP
- Health expenditures as a percentage of GDP
- Inflation rate
- Gini index
- Total internet users

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)



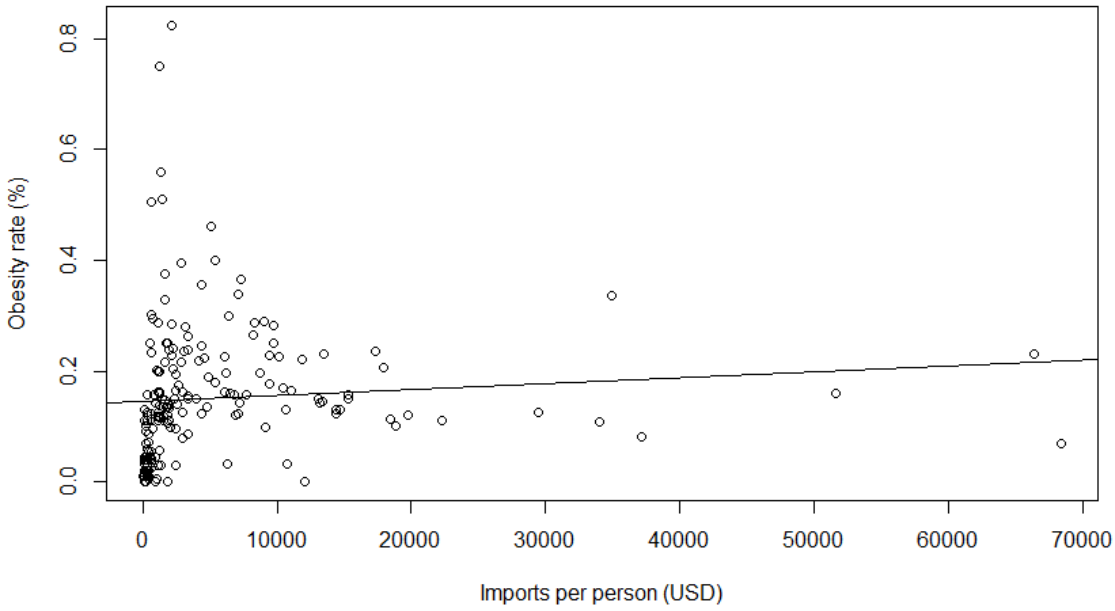
Finally, some variables showed a stronger linear relationship following a transformation of some kind. This group of variables is also listed along with their transformations and their T-test p-values.

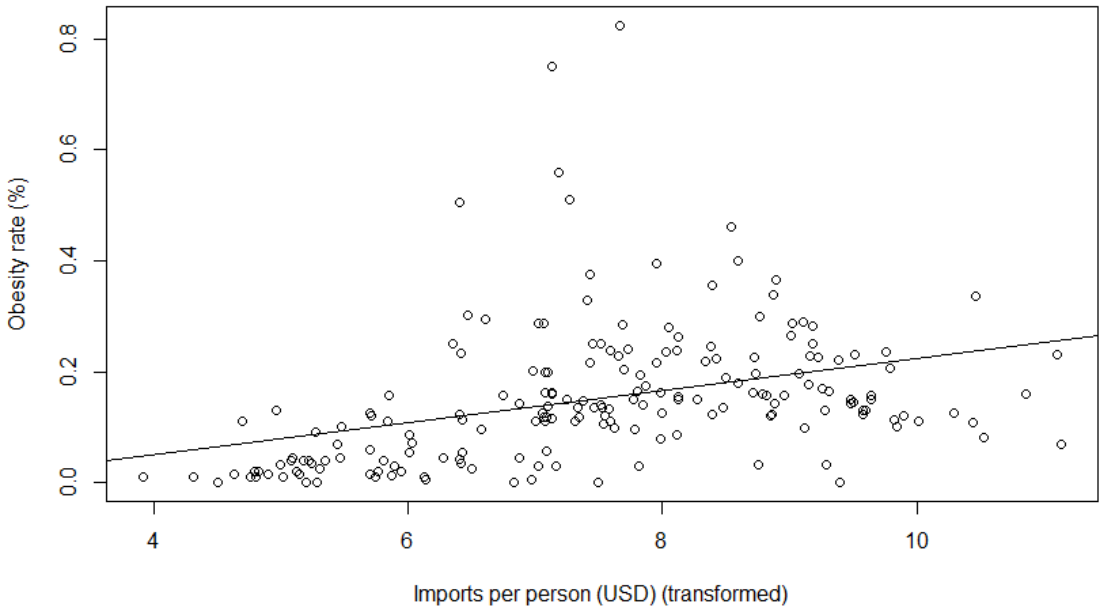
- Per person imports (USD)
  - Transformation: $f(x) = \log x$
  - $p = 2.70 \cdot 10^{-07}$
- Per person GDP (USD)
  - Transformation: $f(x) = \log x$
  - $p = 1.85 \cdot 10^{-07}$
- Percentage of children under age 5 who are underweight
  - Transformation: $f(x) = \sqrt{x}$
  - $p = 2.00 \cdot 10^{-16}$

Choice of graphs and graphical independent variables

The following variables were extracted from our graphs in Python :

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





- From the undirected graph of border relationships:
  - Shortest path to the United States in terms of number of edges
  - Shortest path to the United States in terms of sum of edge weights, defined by border length in kilometers
- From the directed graph of import relationships:
  - Number of incoming edges (i.e., the number of countries which have that country as a major source of imports)
  - TO BE CONTINUED
- From the directed graph of export relationships:
  - Number of outgoing edges (i.e., the number of countries which import a major amount from that country)
  - TO BE CONTINUED

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. There was only one correlation value which was of concern. 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 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.

Validation of usage of graphical variables

ANOVA test used to show that the model with the graphical variables is more informative than the model without them (TO BE CONTINUED)

Figure 4.4: Scatterplot of fertility versus birth rate, overlaid with a simple regression line

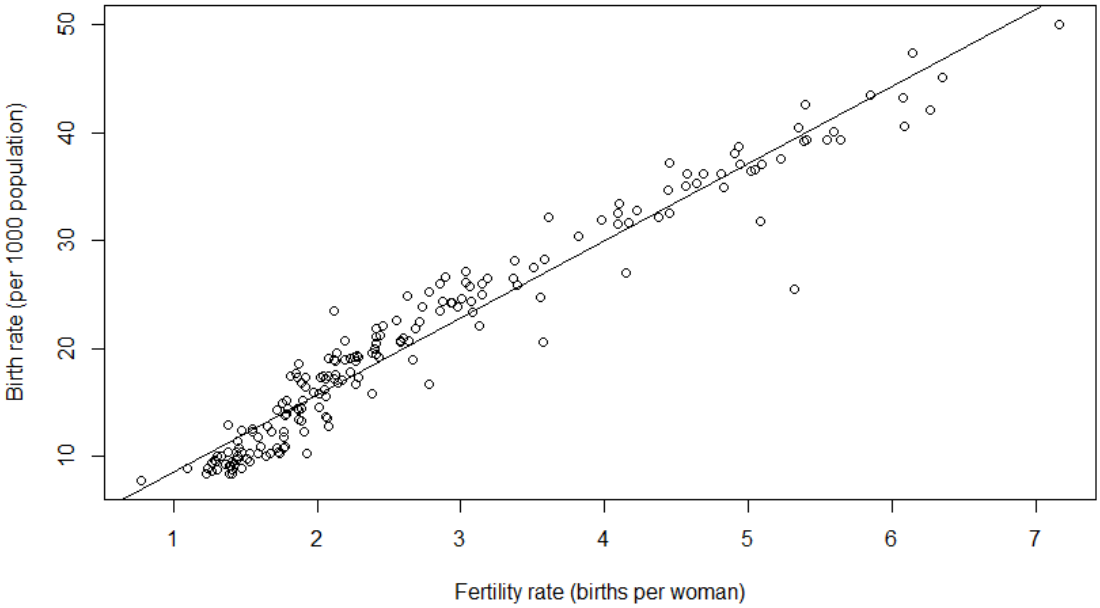
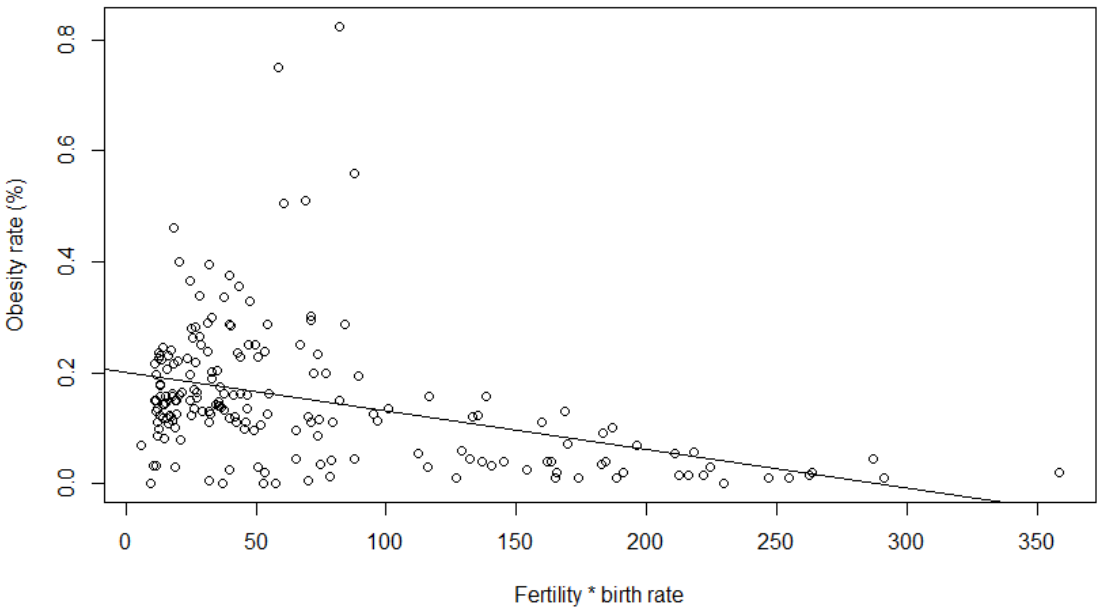


Figure 4.5: Scatterplot of the new variable versus obesity, overlaid with a simple regression line



Chapter 5

Results

Appendix A

Glossary

Obesity. A medical condition that Body Mass Index is greater than 30.

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

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.

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..

Clique. In graph theory, a clique is a subset of a graph's vertices such that every two vertices in the subset are connected by an edge.

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.

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.

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

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.