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 November 5, 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

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

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 is the same thing as South Korea). In order to merge the lists property despite these naming differences, another table was used which listed possible alternate names for each country, and this table was referenced during the merging process to ensure that South Korea and the Republic of Korea had their data combined. As a convention, in the data frame that was built, the name used by the CIA World Factbook tables is the country name listed in the data frame.

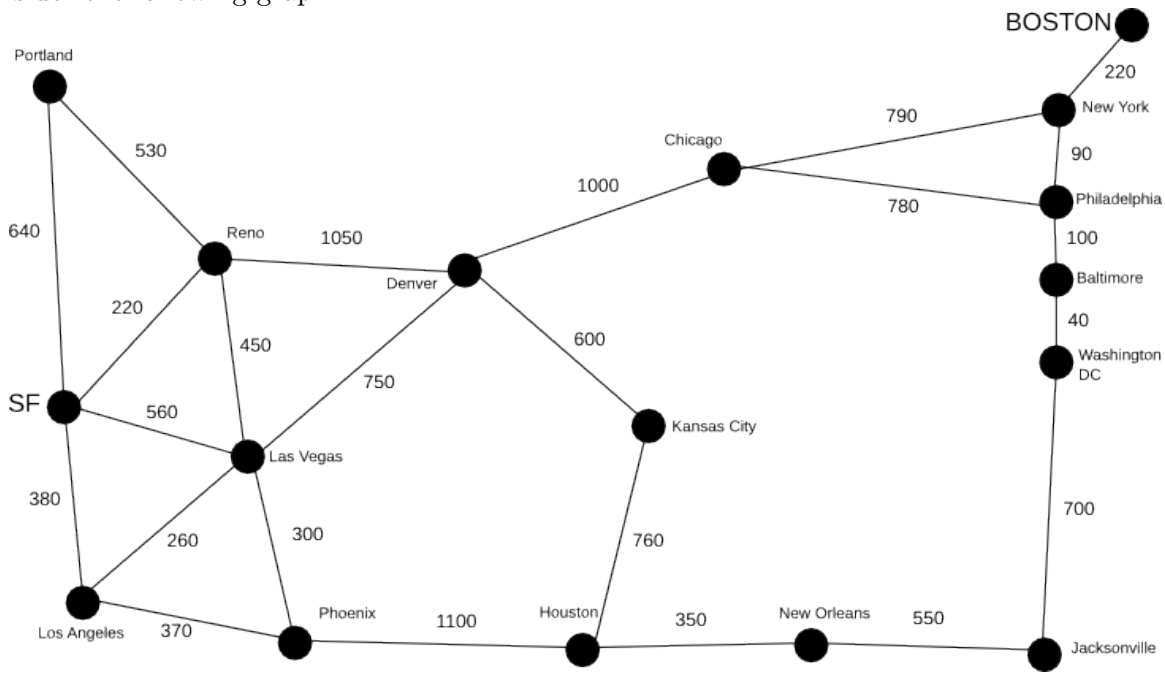
The R environment is effective for statistics and data management, but does not have a robust package for creating graphs. Python, however, does have such a feature. Therefore, the data from R must be converted into a .csv file again. However, individual data points for each country do not allow us to define the vertex-edge graphs that we would like to create. We also need data regarding the relationships between countries, such as bordering countries - this data will be used to define adjacencies in our graphs. This information is not available in a format that is as convenient. It must be mined from a webpage and processed to remove commas, footnotes, and other extraneous text that the CIA has included on the webpage. This was done in Microsoft Excel by pasting the contents of the webpage into a spreadsheet and processing the cells using Excel formulas. Then, the spreadsheet was also exported as a .csv file. Finally, the two .csv files will be imported into Python and used to create the graphs. This step has yet to be completed.

Python allows data to be extracted from graph objects, such as the neighborhood size for each vertex, shortest path data, or clique numbers. These data can also be arranged by country, and added back into the data frame as additional independent variables. After this step, we will have a long list of independent variables in the data frame. Using these variables as predictors, we can then start to build regression models using R's robust regression abilities. It is unlikely that a basic multiple linear regression model will fit the data well. We may need to experiment with other forms of regression, such as exponential or polynomial regression; furthermore, we may need to apply functions like logarithms to the predictor variables so that their relationship to the response variable (obesity rate) follows the assumptions of the model in question. This will be done on an ad-hoc basis based on diagnostics performed on the models, and at this time we cannot predict exactly what measures will be taken in order to build an effective regression model.

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

As implied in the technical background section, the proposed project consists of a few major components: acquisition and structuring of data; graph modeling; regression modeling; and analysis of these results.

The previously-described process for gathering results will leave us with one or more regression models which accurately predict a country's obesity rate. The next questions are: How do we interpret these models? How can the models be used in the future? We said the model will be accurate, but what exactly does that mean?

There exist standard techniques for evaluating the quality of a regression model. We are looking for the residuals to follow a normal distribution; for the slope estimates to have small variances; for an overall hypothesis test of the model's fit to reject a null hypothesis; and for other assumptions, germane to the specific type of regression model, to be met. If our assessments of these factors show that the model fits the data well, then we can validate its use for interpreting the effects of our independent variables on obesity rates.

Ideally, our regression models can be used to experiment with independent variables to predict the effect of changes in these variables on the obesity rate. For example, a policymaker may want to know how he or she can expect the obesity rate to change if health expenditures increase, or if the country imports food products from fewer countries. A model that is accurate by the above standards can give a sound estimate of these effects.

This experimentation will be made possible by a package that we build in R. This package will contain sample data that was used in our building of the models, functions to facilitate manipulation of the independent variables, and functions to create plots and charts of these effects.

Chapter 5

Progress

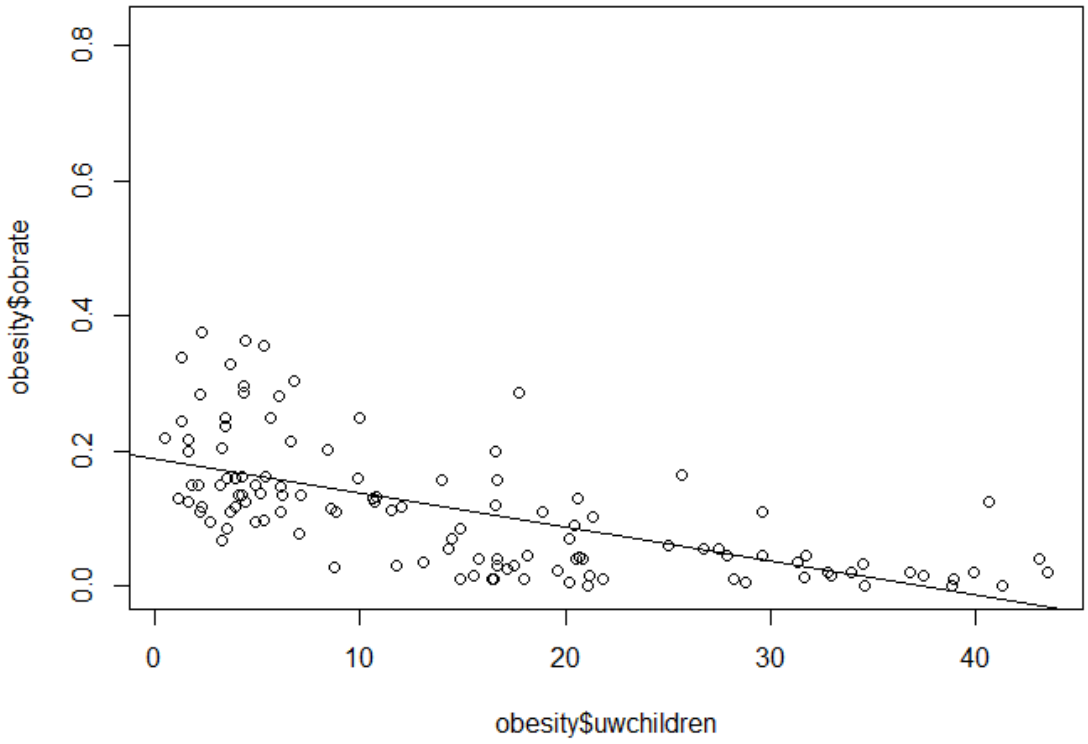
Our work on the project so far has focused on data collection and processing (much of which is described in the technical background section of this report). Data for 190 countries has been collected, with obesity rates as well as 14 other independent variables. These independent variables are as follows:

1. GDP per capita
2. Percentage of children under the age of 5 who are underweight
3. Education expenditures as a percentage of GDP
4. Inflation rate
5. Gini index for families
6. Population growth rate
7. Number of total internet users
8. Health expenditures as a percentage of GDP
9. Birth rate
10. Total electricity consumption
11. Fertility rate
12. Unemployment rate for persons aged 18-24
13. Total population
14. Imports (in US dollars) per person

Below is a snapshot of the data frame in RStudio (a front-end development environment for R).

	X	Country	obrate	ppgdp	uwchildren	education_xp	inflation	gini_index	pop_growth	internet_users	health_xp
1	1	Afghanistan	0.014830918	1000	32.9	NA	7.7	NA	2.22	1000000	7.4
2	2	Albania	0.215380711	7800	6.6	NA	3.5	34.5	0.28	1300000	6.9
3	3	Algeria	0.109504950	7400	3.7	4.3	4.5	35.3	1.17	4700000	5.8
4	4	Andorra	0.236666667	37200	NA	3.2	1.6	NA	0.27	67100	7.7
5	5	Angola	0.055532995	6000	27.5	2.6	13.5	NA	2.78	606700	4.6
6	6	Antigua and Barbuda	0.188421053	18200	NA	2.7	3.3	NA	1.28	65000	5.1
7	7	Argentina	0.375102041	17700	2.3	4.9	22.0	45.8	1.00	13694000	9.5
8	8	Armenia	0.162780749	5500	4.2	3.0	7.7	30.9	0.11	208200	4.7

In an exploratory multiple linear regression model, 6 of these independent variables were significant predictors of obesity rate (according to hypothesis tests on the estimated slope coefficients produced by R): GDP per capita; percentage of underweight children; number of internet users; electricity consumption; total population; and per person imports. The following is a graph of percentages of underweight children versus obesity rate, with a regression line plotted onto it. This relationship was highly statistically significant. Here is also displayed a snapshot of the multiple linear regression model with coefficients and significance markers.



```
> exploratory_model <- lm(obesity[,3:17])
> summary(exploratory_model)

Call:
lm(formula = obesity[, 3:17])

Residuals:
      Min        1Q    Median        3Q       Max
-0.111321 -0.038127 -0.007858  0.028484  0.132400

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.913e-02  7.883e-02   0.877  0.38684
ppgdp               7.229e-06  3.099e-06   2.333  0.02590 *
uwchildren         -6.012e-03  1.649e-03  -3.645  0.00091 ***
education_xp       -1.008e-03  8.180e-03  -0.123  0.90270
inflation           1.029e-03  1.881e-03   0.547  0.58819
gini_index          4.334e-04  1.129e-03   0.384  0.70361
pop_growth          3.452e-02  3.030e-02   1.140  0.26269
internet_users     -3.168e-09  1.250e-09  -2.534  0.01621 *
health_xp           6.898e-03  4.687e-03   1.472  0.15055
birth_rate          5.464e-03  7.411e-03   0.737  0.46618
elec_consum         1.430e-13  7.211e-14   1.983  0.05575 .
fertility          -4.533e-02  3.932e-02  -1.153  0.25730
youth_unemployment  5.425e-05  1.010e-03   0.054  0.95750
population          1.756e-10  8.740e-11   2.009  0.05281 .
ppimports          -8.170e-06  2.841e-06  -2.876  0.00700 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0648 on 33 degrees of freedom
  (142 observations deleted due to missingness)
Multiple R-squared: 0.6796, Adjusted R-squared: 0.5437
F-statistic: 5.001 on 14 and 33 DF,  p-value: 7.071e-05
```

We have also gathered formatted data of all geographic adjacencies between countries.

Going forward, we will start to build graphs in Python and explore ways to infer data from them. The 6 variables which were significant in the exploratory regression tests will be given higher consideration when building the graphs. For example, one piece of data we might consider is the sum of internet users of all of a country's geographical neighbors. There will be many ways to gain information from the graphs, and a lot of the work on this stage of the project will simply be experimenting with different details to find out which information is of interest.

After we have thoroughly perused the graphs in Python, the data we gain from them (e.g. number of neighboring internet users) will be added to the data frame in R, and we can continue with the regression process.

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

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.