A set of graphical models to explore the effects of globalization and development on the prevalence of obesity

Shannon Cebron, Zhendan Zhu, Michael Weinberger

Johns Hopkins University

scebron@cis.jhu.edu

November 27, 2012

Summary

Project overview

-Under-explored phenomenon
-Graphical models to represent global connections
-Statistical models to analyze graphical features

Sponsor

The RAND Corporation

-Nonpartisan public policy research organization
-A focus area is health and health care
-Has previously worked with the US Congress

Background

What is obesity?

-Defined as a BMI greater than 30 (kilograms of body weight per meters of height squared)
-Indicator for heart disease, diabetes, and more
-Causes bone density and joint problems and difficulty with everyday tasks

Background

Costs to society

-Leading cause of preventable death

-Drives up health care costs

-Lost hours affect the economy (117 billion USD per year in the USA)

Background

Relationship to globalization

-Increasing at highest rates in developed countries
-Increasing at higher rates in neighbors of developed countries
-Burger King in Accra, Ghana
-Exploding urbanization in places like Burundi

Deliverables

Regression model

-A model to predict current obesity rate based on known data

Deliverables

Interactive experience

-An R package to encapsulate obesity prediction models
-Allows user manipulation of data to predict effects of policy changes
-Includes sample data

Project outline

Data acquisition

-Obesity rate
-Development status
-Bordering nations
-Economic data
-Trading data
Source: CIA World Factbook

Project outline

Graph construction

-Nations as vertices

-Varying adjacency and edge length definitions for each graph

-Example: Border relationships as adjacencies and distance between capital cities as edge lengths

Environment: Python

Project outline

Data extraction

-Neighborhood sizes

-Shortest path to a developed country

Project outline

Data analysis

Independent variables (examples)

- GDP per capita
- Education expenditures as a percentage of GDP
- Total internet users
- Number of food import partners
- Degree of separation from a highly developed nation
- Rate of urbanization

Dependent variable: Obesity rate

Project outline

Data analysis

-R has great capability for regression diagnostics
-Will consider not only linear, but also exponential, polynomial, and other regressions
-An acceptable model will have a significant F-statistic and significant T-tests for slope coefficients
-Model must also meet assumptions of regression

Preliminary work: data analysis

Figure : Visualization of these data in RStudio

Country	obrate	ppgdp	uwchildren	education_xp	inflation
Afghanistan	0.014830918	1000	32.9	NA	7.7
Albania	0.215380711	7800	6.6	NA	3.5
Algeria	0.109504950	7400	3.7	4.3	4.5
Andorra	0.236666667	37200	NA	3.2	1.6
Angola	0.055532995	6000	27.5	2.6	13.5
Antigua and Barbuda	0.188421053	18200	NA	2.7	3.3
Argentina	0.375102041	17700	2.3	4.9	22.0
Armenia	0.162780749	5500	4.2	3.0	7.7
Australia	0.164000000	40800	NA	4.5	3.4
Austria	0.110000000	42400	NA	5.4	3.5
Azerbaijan	0.201020408	10300	8.4	2.8	8.1
Bahamas, The	0.290000000	31400	NA	NA	3.2
Bahrain	0.282340426	27900	NA	2.9	-0.4
Bangladesh	0.000000000	1700	41.3	2.4	10.7

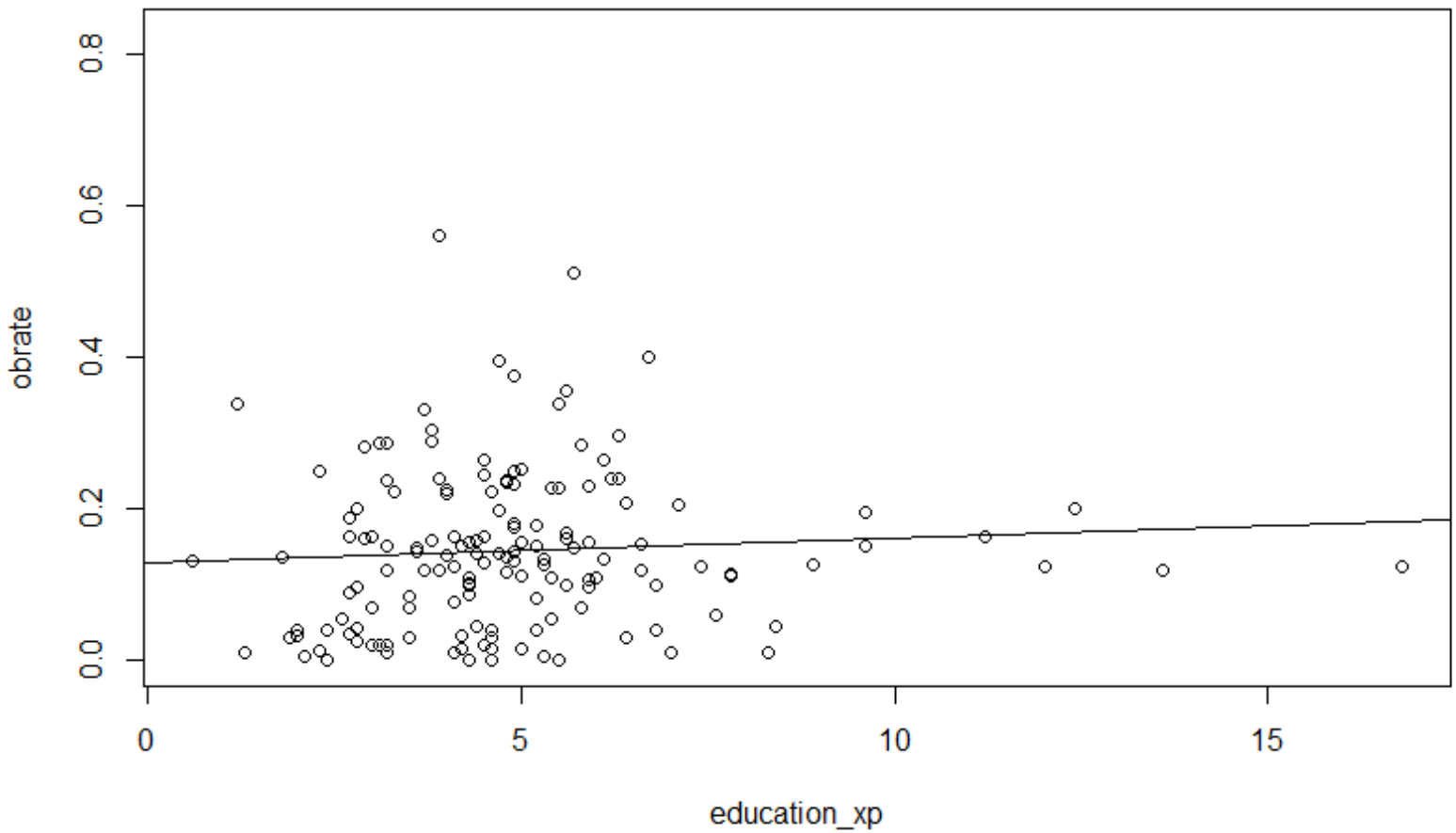
Preliminary work: data analysis

Figure : Simple regression of obesity rate on education expenditures

```
Call:
lm(formula = obrate ~ education_xp, data = obesity)

Residuals:
     Min       1Q   Median       3Q      Max
-0.14620 -0.08373 -0.01251  0.05931  0.41896

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.128476   0.020071   6.401 1.87e-09 ***
education_xp 0.003222   0.003750   0.859    0.392
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1046 on 150 degrees of freedom
  (38 observations deleted due to missingness)
Multiple R-squared: 0.004897, Adjusted R-squared: -0.001737
F-statistic: 0.7381 on 1 and 150 DF,  p-value: 0.3916
```

Preliminary work: data analysis

Figure : Plot of obesity rate against education expenditures, overlaid with regression line



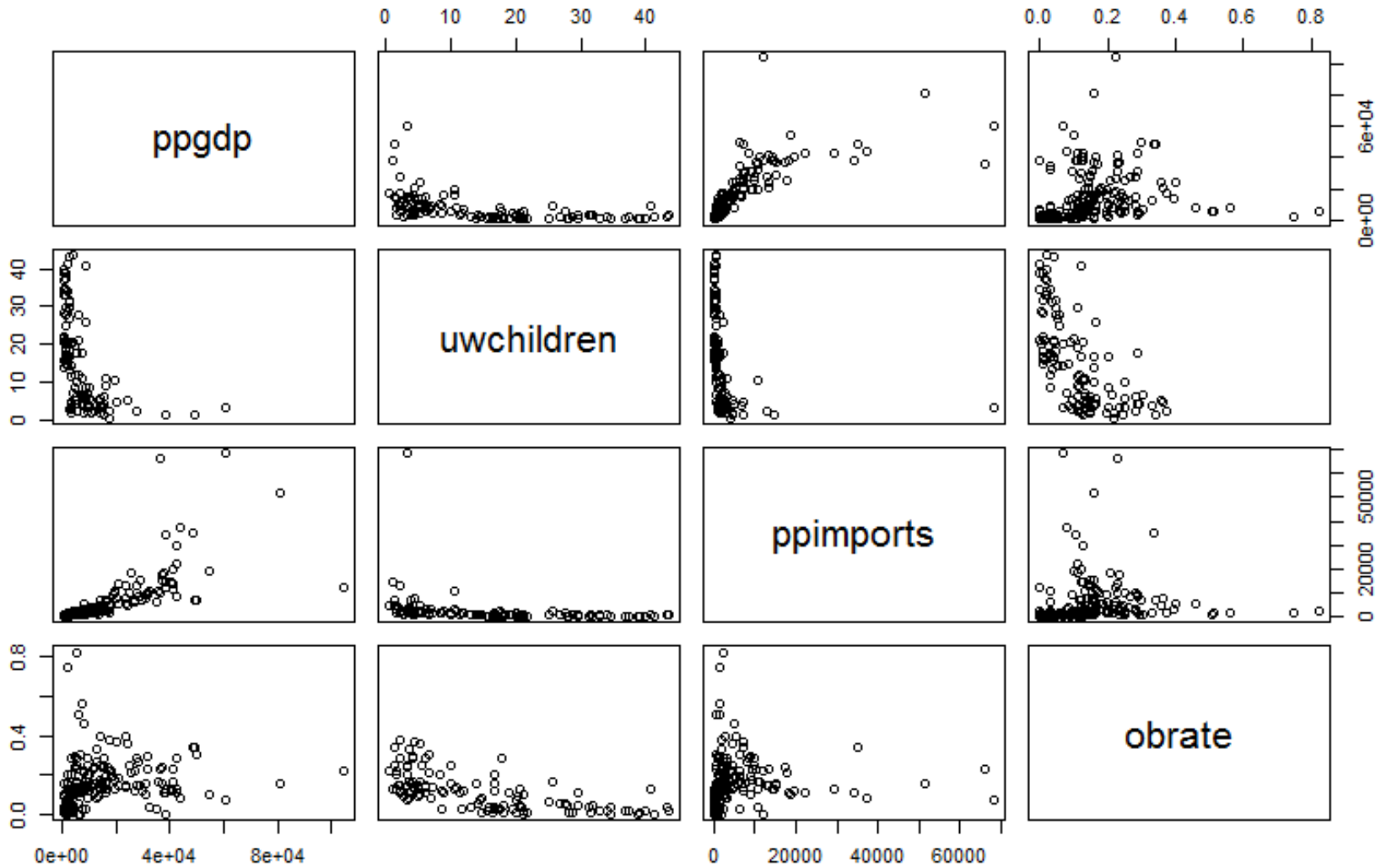
Preliminary work: data analysis

Figure : Multiple linear regression of obesity rate against basic independent variables

```
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
```

Preliminary work: data analysis

Figure : Scatterplox matrix of several relationships



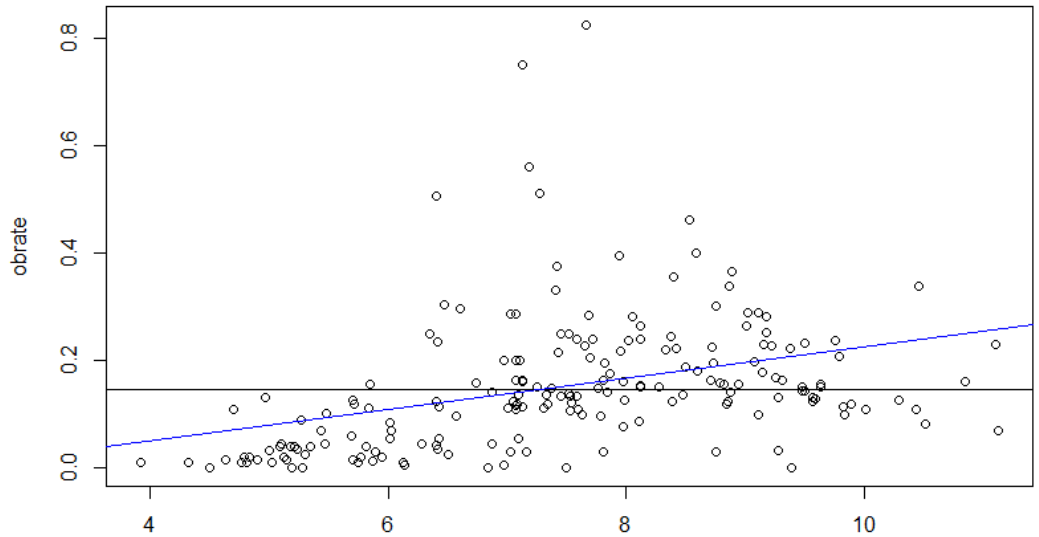
Preliminary work: data analysis

Discovering nonlinear relationships

-Per person imports has a weak linear relationship with obesity rate (p=0.269)

-Log per person imports has a strong linear relationship with obesity rate (p=0.00000027)

Figure : Log(imports) takes on a stronger relationship with the data



Preliminary work: graph construction

Graph constructed in Python showing national border relationships

Conclusion

Remaining work to be done

-Extracting data from graphs and importing into R
-Regression analysis using this new data
-Creation of deliverables

Looking forward

-Future research might including a more in-depth exploration of graphical variables
-Efforts may be taken to find obesity data for more than 190 countries
-More lifestyle data and cultural factors could be considered, such as popularity of sports

The End