Quantifying Correlations and Causal Inference

Anav Vora, Hari Dave, Yi-Chia Chang
Nov. 17, 2023

Introduction

We examine the relationship between the betweenness centrality of a country and the strength of virtual water trade corresponding to that country. Countries that lie on a large number of paths connecting other countries in a network have a higher betweenness centrality. Strength is defined as the total volume of virtual water that a country imports or exports. Countries that have a higher betweenness may be important donors of food aid or recipients of large quantities of food aid. Consequently, large volumes of virtual water may be imported or exported from countries with higher betweenness centrality, thus increasing the strength of that country. Before examining the relationship, we take the logarithm of strength and betweenness to induce normality in both variables. A linear relationship is noted between log(Betweenness) and log(Strength), suggesting that the hypothesized relationship may be true (Figure 1).

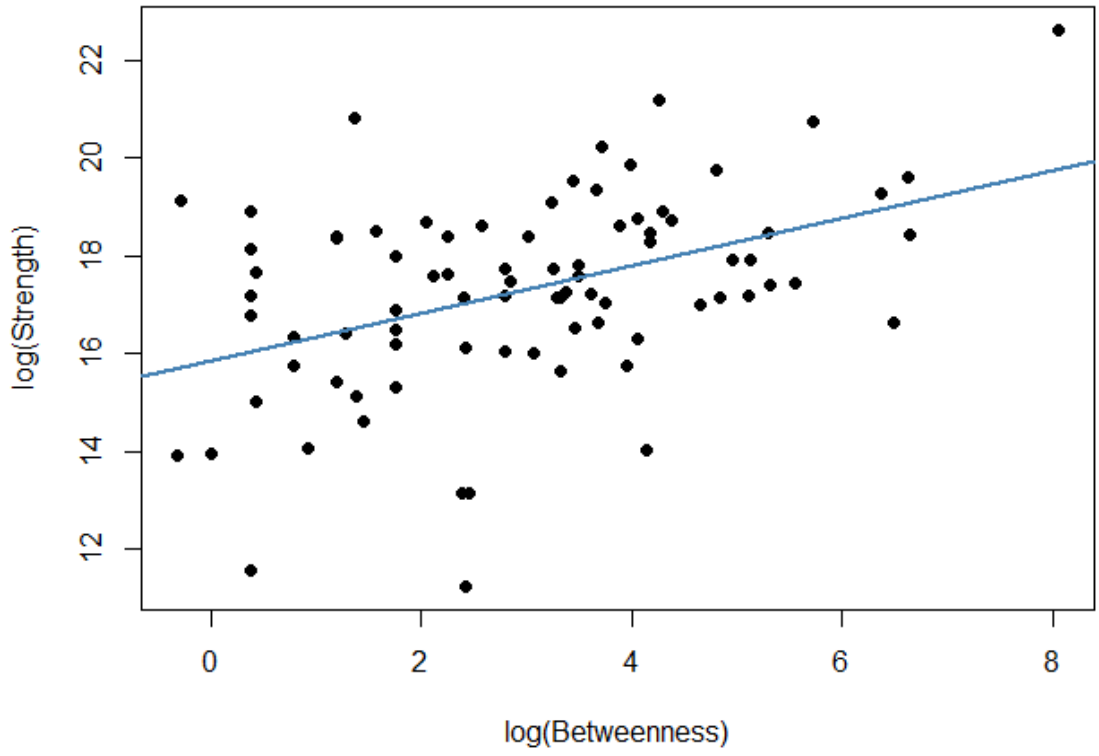

Figure 1: Scatterplot showing the linear positive relationship between log(Betweenness) and log(Strength).

Ordinary least squares (OLS) regression

Constructing an OLS regression line (shown in blue in Figure 1) demonstrates a statistically significant relationship between log(Betweenness) and log(Strength) with a p-value of 3.4e-05 (i.e., <0.05) (Table 1).

Table 1: Coefficients for the regression equation between the predictand log(Strength) and the predictor log(Betweenness). The significance is indicated using *** to demonstrate a p-value less than 0.05. F-statistics is 19.26.

	Estimate	Std. Error	p-value
Intercept	15.84	0.39	< 2e-16***
log(Betweenness)	0.49	0.11	3.4e-05***

Treatment in the relationship

The “treatment” in the relationship for this case is the betweenness centrality, and the outcome variable is the strength of virtual water trade.
The gold standard for causal impact inference is randomized control trials (RCT) for dealing with selection bias. However, it is inapplicable for our case as we cannot induce a group of countries to increase trade, thereby increasing their betweenness centrality. Furthermore, in a network, the betweenness centrality of a country can change due to changes in the trade relationships of the countries it trades with.

Issues with inferring causality from the relationship modeled using OLS

We recognize that the economic and climatic conditions of a country can impact the strength of virtual water trade. For example, two countries may have the same betweenness centrality but have different values of strength due to their economic conditions. A richer country might export relatively larger quantities of food aid (or virtual water) than a relatively poorer country with the same betweenness centrality. Similarly, two countries may have the same betweenness centrality, but one of them may be a large donor, while the other may be a recipient in most of its trade relationships. It would be unreasonable to expect the same strength of virtual water trade for both countries in such a case. In addition, variables such as the economy of the country, or the income of a nation can affect both the centrality and the strength of the trade. Richer countries can facilitate trade and, thus, can have higher centrality. In addition, they can export more goods and therefore have a greater virtual water trade. Due to these reasons, causality cannot be directly inferred from the relationship noted via OLS regression.

Econometric approach to evaluate the causal relationship

The instrumental variable approach can be used to evaluate the causality of the relationship noted in Figure 1. For the instrumental variable technique, we must identify a variable that is correlated with the independent variable (betweenness in our case), but not with the dependent

variable (strength in our case). Mean distance to other countries may be a good instrumental variable in our case, as countries located close to most other countries may serve as good transportation hubs, and consequently have a higher betweenness centrality. However, there shouldn't be a theoretical relationship between distance and strength of virtual water trade (other than the relationship through betweenness centrality).

To apply the instrumental variable technique, we must first develop a regression relationship between mean distance to other countries and the betweenness centrality. Next, predicted betweenness from this regression relationship should be used as the independent variable for a second regression relationship (between constructed betweenness and strength). Coefficients and statistical significance of the second relationship must then be examined to determine causality. In this way, the instrumental variable method is superior to the OLS regression technique as it would allow us to examine causality that may be impacted by issues such as omitted variables, endogeneity, and selection bias.

Appendix

```
rm(list=ls())
library(igraph)
library(network)
library(qgraph)
Data=read.csv("./AWF_agg.csv")

#No of nodes: is the number of unique countries participating in trade
Nodes = unique(c(unique(Data$country_export),unique(Data$country_import)))
NoOfNodes = length(Nodes)
#Strength (Volume of water going in or coming out of a node)
Strength_In  = 0; Strength_Out = 0; Strength_Tot = 0
#Analysing for each node using a for loop
for(i in 1:NoOfNodes){
  #Extracting a subset of the main data containing countries
  #receiving virtual water from the selected node
  Subset_Out = Data[which(Data$country_export==Nodes[i]),]
  #Extracting a subset of the main data containing countries
  #sending virtual water to the selected node
  Subset_In  = Data[which(Data$country_import==Nodes[i]),]
  #Summing over the volume of virtual water received by a node
  Strength_In[i] = sum(Subset_In$Actual.WF..m3.)
  #Summing over the volume of virtual water sent by a node
  Strength_Out[i] = sum(Subset_Out$Actual.WF..m3.)
  #Total strength is the summation of received and sent virtual water
  Strength_Tot[i] = Strength_In[i] + Strength_Out[i]
}
OutputMatrix = data.frame(Country = Nodes,Strength_In,Strength_Out,Strength_Tot)
#Converting the data of trades into a graph
#The output is a list with 113 elements (I.e., one element for each node)
#The value in the list is the nodes that that node is connected to
Data_Graph = graph.data.frame(Data[,c(2:4)])
#Getting the betweeness centrality for each node
Centrality = betweenness(Data_Graph,directed = F)

#Plotting betweeness against total strength
For_Regression = data.frame(y=log(Strength_Tot),x=log(Centrality))
For_Regression$y[which(For_Regression$y==-Inf)] = NA
For_Regression$x[which(For_Regression$x==-Inf)] = NA
For_Regression = na.omit(For_Regression)
#Showing normality of variables
qqnorm(For_Regression$y)
qqline(For_Regression$y)
qqnorm(For_Regression$x)
qqline(For_Regression$x)
#Plotting scatterplot
plot(For_Regression$x,For_Regression$y,xlab="log(Betweenness)",ylab="log(Strength)",pch=16)
#Creating a regression between log(Centrality) and log(Strength_tot)
model <- lm(y ~ x, data = For_Regression)
#Adding regression line to scatterplot
abline(model,col="steelblue",lwd=2)
#Checking if the OLS regression is statistically significant
summary(model)
```

All codes and figures can be found in the GitHub repository:
https://github.com/yichiac/CEE598_Globalization_of_Water/hw3