ECO7707 - Problem Set I

Armand Kapllani

1/14/2018

1. Data Collection and Gravity Equation

This homework asks you to collect trade data in R and estimate the gravity equation. In particular it asks you to evaluate the impact of the Euro on trade flows using both an OLS as well as a diff-in-diff approach. Finally, it asks you to discuss potential threats to identification.

1.1. Use the function get.Comtrade() which we discussed in class to download all trade flows between the 20 largest countries for the year 2014.1 Collect only information on the total trade (i.e. use c = "TOTAL") but get information on both imports and exports.

Clean the global environment in R.

```
rm(list = ls())
```

Install/upload required packages.

```
library(data.table)
library(ggplot2)
library(stargazer)
library(rjson)
```

The following function downloads and formats the data.

```
get.Comtrade <- function(
  # construct the url for downloading
  url="http://comtrade.un.org/api/get?"
  ,maxrec=10000
  ,type="C"
  ,freq="A"
  ,px="HS"
  ,ps="now"
  ,r
  ,p
  ,rg="all"
  ,cc="TOTAL"
  ,fmt="json"
)
{
  string<- paste(url
                 ,"max=",maxrec,"&" # maximum no. of records returned
                 ,"type=",type,"&"  # type of trade (c=commodities)
                 ,"freq=",freq,"&"  # frequency
                 ,"px=",px,"&"      # classification
                 ,"ps=",ps,"&"      # time period
                 ,"r=",r,"&"        # reporting area
                 ,"p=",p,"&"        # partner country
                 ,"rg=",rg,"&"      # trade flow
```

```
                  ,"cc=",cc,"&"      # classification code
                  ,"fmt=",fmt        # Format
                  ,sep = ""
  )
  raw.data<- fromJSON(file=string)
  data<- raw.data$dataset
  print(data)
  validation<- unlist(raw.data$validation, recursive=TRUE)
  var.names<- names(data[[1]])
  data<- as.data.frame(t( sapply(data,rbind)))
  ndata<- NULL
  for(i in 1:ncol(data)){
    data[sapply(data[,i],is.null),i]<- NA
    ndata<- cbind(ndata, unlist(data[,i]))
  }
  ndata<- as.data.frame(ndata)
  colnames(ndata)<- var.names
  return(ndata)
}
```

Import the datasets on countries GDP for two years 2014 and 1995 and distance between each pair.

```
gdp_1995 <- read.csv("/Users/armandkapllani/Desktop/UF Econ/International
                     Economic Relations/PS1/gdp_1995.csv", header = T)
gdp_2014 <- read.csv("/Users/armandkapllani/Desktop/UF Econ/International
                     Economic Relations/PS1/gdp_2014.csv", header = T)
distance <- read.csv("/Users/armandkapllani/Desktop/UF Econ/International
                     Economic Relations/PS1/distance.csv", header = T)

gdp_1995 <- data.table(gdp_1995)
gdp_2014 <- data.table(gdp_2014)
```

Use the following loops in combination with Comtrade() function to obtain all trade flows between pairs for all the twenty largest countries.

```
I_C <- as.character(c(gdp_2014$cty_code))
d1 <- NULL
for (i_c in I_C[1:5]) {
  for (i_k in I_C) {
    if (i_c == i_k) {
      next
    }
    else if (i_c != i_k) {
      data_i_c <- get.Comtrade(r = i_c, p = i_k, rg = "2", c = "TOTAL", ps = "2014")
      d1 <- rbind(d1, data_i_c)
    }
  }
  Sys.sleep(10)
}

Sys.sleep(3660)

I_C <- as.character(c(gdp_2014$cty_code))
d2 <- NULL
for (i_c in I_C[6:10]) {
```

```
  for (i_k in I_C) {
    if (i_c == i_k) {
      next
    }
    else if (i_c != i_k) {
      data_i_c <- get.Comtrade(r = i_c, p = i_k, rg = "2", c = "TOTAL", ps = "2014")
      d2 <- rbind(d2, data_i_c)
    }
  }
  Sys.sleep(10)
}

Sys.sleep(3660)

I_C <- as.character(c(gdp_2014$cty_code))
d3 <- NULL
for (i_c in I_C[11:15]) {
  for (i_k in I_C) {
    if (i_c == i_k) {
      next
    }
    else if (i_c != i_k) {
      data_i_c <- get.Comtrade(r = i_c, p = i_k, rg = "2", c = "TOTAL", ps = "2014")
      d3 <- rbind(d3, data_i_c)
    }
  }
  Sys.sleep(10)
}

Sys.sleep(3660)

I_C <- as.character(c(gdp_2014$cty_code))
d4 <- NULL
for (i_c in I_C[16:20]) {
  for (i_k in I_C) {
    if (i_c == i_k) {
      next
    }
    else if (i_c != i_k) {
      data_i_c <- get.Comtrade(r = i_c, p = i_k, rg = "2", c = "TOTAL", ps = "2014")
      d4<- rbind(d4, data_i_c)
    }
  }
  Sys.sleep(10)
}

dta <- rbind(d1, d2, d3, d4)
dta <- data.table(data)

  ## Imports (2014)

I_C <- as.character(c(gdp_2014$cty_code))
d5 <- NULL
```

```
for (i_c in I_C[1:5]) {
  for (i_k in I_C) {
    if (i_c == i_k) {
      next
    }
    else if (i_c != i_k) {
      data_i_c <- get.Comtrade(r = i_c, p = i_k, rg = "1", c = "TOTAL", ps = "2014")
      d5 <- rbind(d5, data_i_c)
    }
  }
  Sys.sleep(10)
}

Sys.sleep(3660)

I_C <- as.character(c(gdp_2014$cty_code))
d6 <- NULL
for (i_c in I_C[6:10]) {
  for (i_k in I_C) {
    if (i_c == i_k) {
      next
    }
    else if (i_c != i_k) {
      data_i_c <- get.Comtrade(r = i_c, p = i_k, rg = "1", c = "TOTAL", ps = "2014")
      d6 <- rbind(d6, data_i_c)
    }
  }
  Sys.sleep(10)
}

Sys.sleep(3660)

I_C <- as.character(c(gdp_2014$cty_code))
d7 <- NULL
for (i_c in I_C[11:15]) {
  for (i_k in I_C) {
    if (i_c == i_k) {
      next
    }
    else if (i_c != i_k) {
      data_i_c <- get.Comtrade(r = i_c, p = i_k, rg = "1", c = "TOTAL", ps = "2014")
      d7 <- rbind(d7, data_i_c)
    }
  }
  Sys.sleep(10)
}

Sys.sleep(3660)

I_C <- as.character(c(gdp_2014$cty_code))
d8 <- NULL
for (i_c in I_C[16:20]) {
  for (i_k in I_C) {
```

```
    if (i_c == i_k) {
      next
    }
    else if (i_c != i_k) {
      data_i_c <- get.Comtrade(r = i_c, p = i_k, rg = "1", c = "TOTAL", ps = "2014")
      d8 <- rbind(d8, data_i_c)
    }
  }
  Sys.sleep(10)
}
```

Use data.table().

```
data_countries_2014 <- rbind(d1, d2, d3, d4, d5, d6, d7, d8)
data_countries_2014 <- data.table(data_countries_2014)
```

1.2. Match each country pair to its bilateral distance using the file "distance.RData". In this file "o" and "d" denote the origin and destination country, respectively. Distance is reported in km between a set of the biggest cities of the two countries.

```
keycols <- c("cty_code_o","cty_code_d")
setnames(data_countries_2014, c("rtCode","ptCode"), keycols)
setnames(gdp_2014, "cty_code", "cty_code_o")
```

Convert column classes to numeric.

```
data_countries_2014$cty_code_o <- as.numeric(as.character(data_countries_2014$cty_code_o))
data_countries_2014$cty_code_d <- as.numeric(as.character(data_countries_2014$cty_code_d))
```

Match by distance. We will the merge() function to perform the matching process.

```
data_countries_2014 <- merge(data_countries_2014, distance, by = keycols)
data_countries_2014$dist <- as.numeric(as.character(data_countries_2014$dist))
```

1.3. Match each importing and exporting country to its respective country GDP in 2014 using the file "gdp 2014.RData".

```
gdp_2014 <- data.table(gdp_2014)
data_countries_2014 <- data.table(data_countries_2014)
```

Match for gdp_2014_i We create the following loop in order to perform the matching.

```
data_countries_2014$gdp_2014_i <- NA
I <- as.numeric(c(gdp_2014$cty_code_o))
K <- as.numeric(c(data_countries_2014$cty_code_o))

for (i in seq_along(I)) {
  for (j in seq_along(K)) {
    if (data_countries_2014$cty_code_o[j] == gdp_2014$cty_code_o[i]) {
      data_countries_2014$gdp_2014_i[j] <- gdp_2014$gdp_2014[i]
      }
    else if (data_countries_2014$cty_code_o[j] != gdp_2014$cty_code_o[i]) {
      next
      }
```

```
  }
}
```

Match for gdp_2014_j We create the following loop in order to perform the matching.

```
data_countries_2014$gdp_2014_j <- NA
I <- as.numeric(c(gdp_2014$cty_code_o))
K <- as.numeric(c(data_countries_2014$cty_code_o))

for (i in seq_along(I)) {
  for (j in seq_along(K)) {
    if (data_countries_2014$cty_code_d[j] == gdp_2014$cty_code_o[i]) {
      data_countries_2014$gdp_2014_j[j] <- gdp_2014$gdp_2014[i]
    }
    else if (data_countries_2014$cty_code_d[j] != gdp_2014$cty_code_o[i]) {
      next
    }
  }
}
```

Now let's keep the variables we need for our analysis.

```
data_countries_2014 <- data_countries_2014[ , .(rgDesc, rtTitle, cty_code_o, ptTitle,
                                                cty_code_d, dist, gdp_2014_i, gdp_2014_j,
                                                TradeValue)]

data_countries_2014 <- data.frame(data_countries_2014)
data_countries_2014$TradeValue <- as.numeric(as.character(data_countries_2014$TradeValue))
data_countries_2014 <- data.table(data_countries_2014)

setnames(data_countries_2014, c("rgDesc", "dist"), c("TradeFlow", "Distance"))
```

Note:

```
    # TradeFlow:  Export or Import.
    # Distance:   Distance measured in km between country i and country j.
    # gdp_2014_i: GDP of country i.
    # gdp_2014_j: GDP of country j.
    # TradeValue: The total value of trade from country i to country j.
```

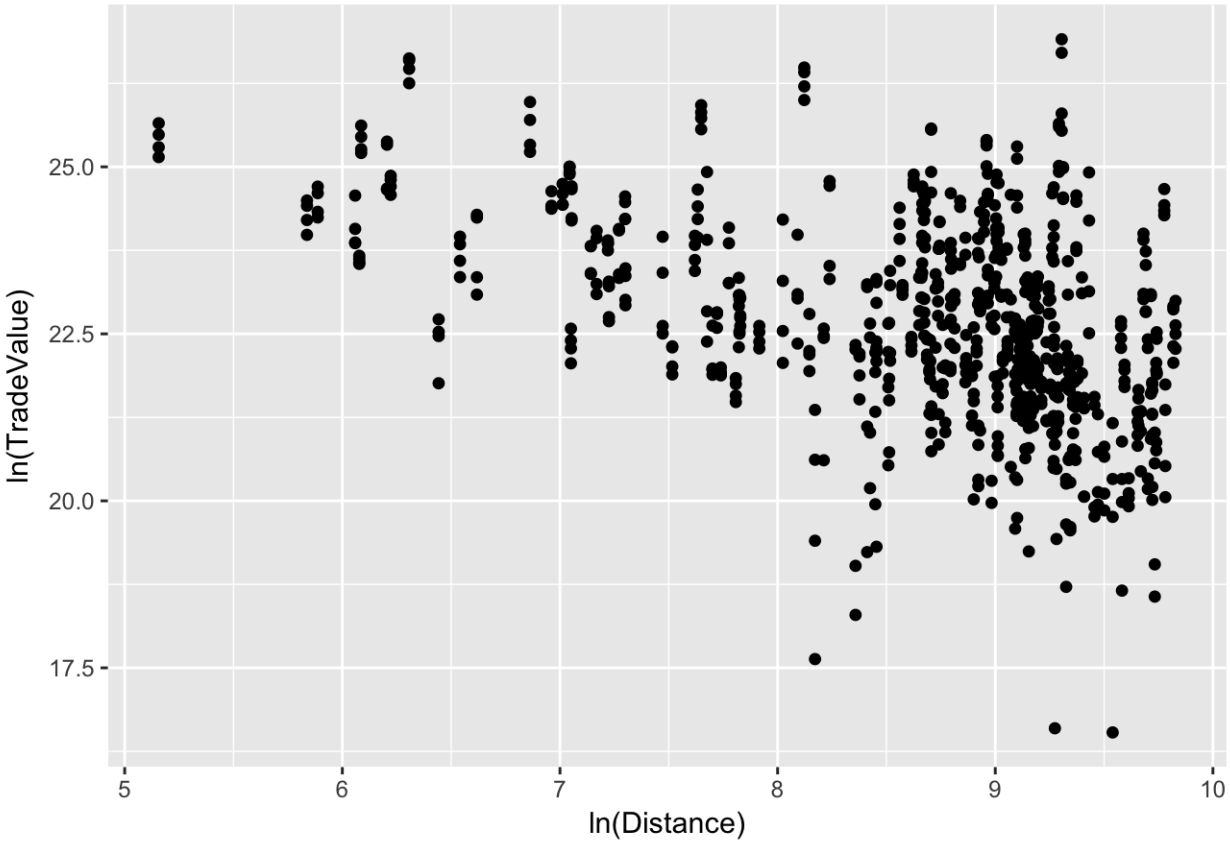
1.4. Using ggplot2, plot the natural log of trade flows on the y-axis and ln(distance) on the x-axis. Is the plot consistent with the gravity equation? Repeat the same exercise but use the natural log of the exporting country's GDP on the x-axis instead.

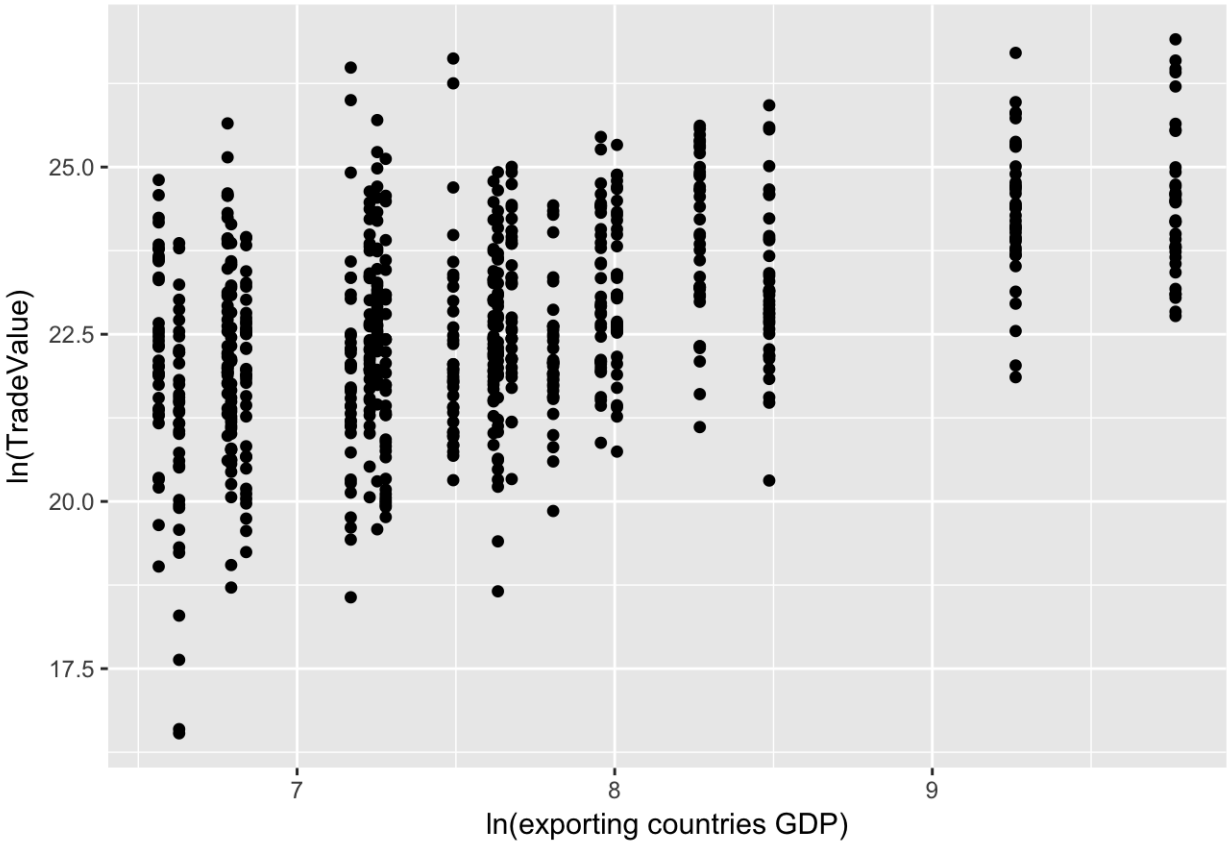
Plot the natural log of trade flows on the y-axis and ln(distance) on the x-axis Plot the natural log of trade flows on the y-axis and log of exporting country's GDP on the x-axis. We make use of the function with(), in this way we don't have to create a new dataset.

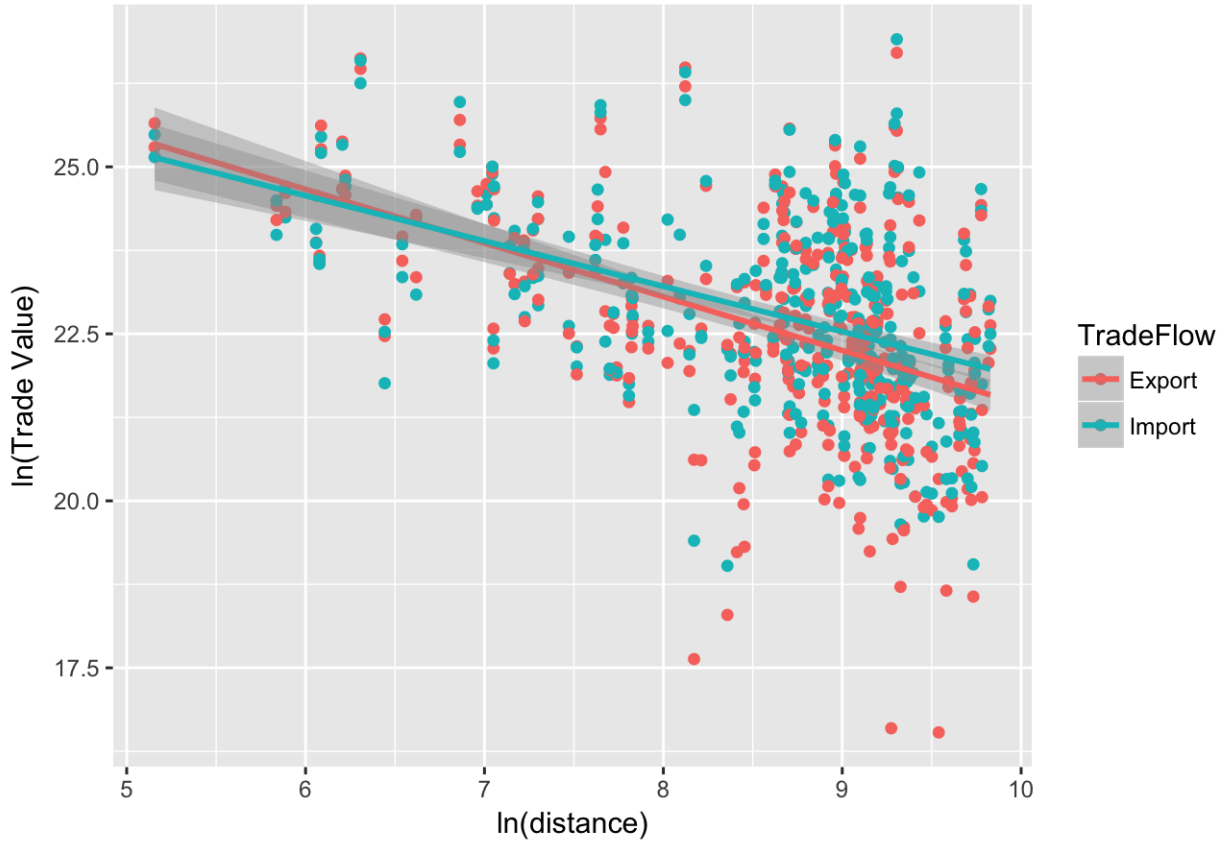
```
load(file = "Untitled.RData")
library(ggplot2)
plot1 <- ggplot(data_countries_2014, aes(log(Distance), log(TradeValue))) +
    geom_point() + xlab("ln(Distance)") + ylab("ln(TradeValue)")
plot1
```

```r
plot2 <- with(data_countries_2014[data_countries_2014$TradeFlow == "Exports"],
    ggplot(data_countries_2014, aes(log(gdp_2014_i), log(TradeValue))) + geom_point()) +
    xlab("ln(exporting countries GDP)") + ylab("ln(TradeValue)")
plot2

plot3 <- ggplot(data_countries_2014, aes(log(Distance), log(TradeValue), colour = TradeFlow)) +
    geom_point() + xlab("ln(distance)") + ylab("ln(Trade Value)") + geom_smooth(method = "lm",
    formula = y ~ x)
plot3
```


ln(TradeValue)
25.0
22.5
20.0
17.5
5
6
7
8
9
10
ln(Distance)


ln(TradeValue)
25.0
22.5
20.0
17.5
7
8
9
ln(exporting countries GDP)



1.5. Run the following gravity equation using the lm() function in R:

$$\ln(x_{ij}) = \beta_0 + \beta_1 \ln(y_i) + \beta_2 \ln(y_j) + \beta_3 \ln(dist_{ij}) + u_{ij}$$

where $x_{i,j}$ denotes a trade flow from country i to country j, y_i and y_j the countries' GDPs and $dist_{ij}$ the

bilateral distance between the two. Further, assume that u_{ij} is a mean-zero error term and uncorrelated with the regressors.

```
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```
grav_eq <- lm(log(TradeValue) ~ log(gdp_2014_i) + log(gdp_2014_j) + log(Distance),
              data = data_countries_2014)

stargazer(grav_eq, title = "Estimation of Gravity Equation",
          covariate.labels = c("$ln(y_i)$", "$ln(y_j)$", "$ln(Distance)$"),
          header = FALSE, digits = 3)
```

Table 1: Estimation of Gravity Equation

	Dependent variable:
	log(TradeValue)
$ln(y_i)$	0.980^{***}
	(0.038)
$ln(y_j)$	0.870^{***}
	(0.038)
$ln(Distance)$	-0.800^{***}
	(0.033)
Constant	15.466^{***}
	(0.500)
Observations	760
R^2	0.681
Adjusted R^2	0.680
Residual Std. Error	0.866 (df = 756)
F Statistic	537.405^{***} (df = 3; 756)
Note:	*p<0.1; **p<0.05; ***p<0.01

1.6. What do you find? Are your results comparable to the ones by Frankel et al (1995) **that we covered in class?**

```
We find that the estimates on ln(country i's GDP) and on ln(country j's GDP) are both positive
and less than one. While the estimate on ln(dist) is negative as expected. Hence the Gravity
equation holds as expected. Yes the results are comparable with what we covered in class. Since
this is a log-log model we would interpret the results as follows:

If country i's GDP increases by 1 % then trade flow will increase by 0.979 %. (more trade)
If country j's GDP increases by 1 % then trade flow will increase by 0.869 %. (more trade)
If the distance between two countries increases by 1% then trade flow will
decrease by 0.799 %. (less trade)
Based on R2, 68 % of the variation in trade flow can be explained by the three covariates
that we controlled for.
```

A common argument for the Euro is that it facilitates and spurs trade. Suppose you want to look into this claim and decide to augment the above gravity equation to do so.

1.7. Create a dummy that is one if both countries use the Euro today and 0 otherwise.

```
## Eurozone countries: Austria, Belgium, Cyprus, Estonia, Finland, France (251),
## Germany (276), Greece, Ireland, Italy (381), Latvia, Lithuania, Luxembourg, Malta,
## the Netherlands (528), Portugal, Slovakia, Slovenia, and Spain (724).

## In our dataset the countries in Eurozone are: France (251), Germany (276), Italy (381),
## the Netherlands (528), Spain (724).
```

Let's create all the possible combinations of the five Eurozone countries using the Eurozone country codes.

```
euro <- c(251, 276, 381, 528, 724)
comb <- data.table(expand.grid(euro, euro))
setnames(comb, c("Var1", "Var2"), c("cty_code_o", "cty_code_d"))
comb <- data.frame(comb)
```

Create a dummy variable using the cty_code_O and cty_code_d as pairs.

```
M <- as.numeric(c(comb$cty_code_o))
L <- as.numeric(c(data_countries_2014$cty_code_o))

EuroDummy <- matrix(NA, nrow(data_countries_2014))

for (j in seq_along(M)) {
  for (i in seq_along(L)) {
    if ( (comb$cty_code_o[j] == data_countries_2014$cty_code_o[i])  &&
         (comb$cty_code_d[j] == data_countries_2014$cty_code_d[i]) )
      EuroDummy[i] <- i
  else
      next
  }
}

data_countries_2014$EuroDummy <- ifelse(is.na(EuroDummy), 0, 1)
```

1.8. Reestimate the gravity equation but also include the Euro dummy, i.e. estimate using OLS as before. I{Euro} is equal to 1 if both i and j use the Euro and 0 otherwise. What do you find regarding the coefficient on this dummy variable? How does the R^2 compare to the one for regression

$$ln(x_{ij}) = \beta_0 + \beta_1 \ln(y_i) + \beta_2 \ln(y_j) + \beta_3 \ln(dist_{ij}) \beta_4 I\{Euro\} + u_{ij}$$

using OLS as before. $IEuro$ is equal to 1 if both i and j use the $Euro$ and 0 otherwise. What do you find regarding the coefficient on this dummy variable? How does the R^2 compare to the one for regression (1)?

```
grav_eq_euro <- lm(log(TradeValue) ~ log(gdp_2014_i) + log(gdp_2014_j) + log(Distance) + EuroDummy,
              data = data_countries_2014)

stargazer(grav_eq_euro,
          title            = "Estimation of Gravity Equation controlling for Eurozone",
          covariate.labels = c("$ln(y_i)$", "$ln(y_j)$", "$ln(Distance)$", "EuroDummy"),
          header = FALSE)
```

Table 2: Estimation of Gravity Equation controlling for Eurozone

	Dependent variable:
	log(TradeValue)
$ln(y_i)$	0.979^{***}
	(0.038)
$ln(y_j)$	0.869^{***}
	(0.038)
$ln(Distance)$	-0.758^{***}
	(0.037)
EuroDummy	0.374^{**}
	(0.159)
Constant	15.097^{***}
	(0.522)
Observations	760
R^2	0.683
Adjusted R^2	0.681
Residual Std. Error	0.864 (df = 755)
F Statistic	406.845^{***} (df = 4; 755)
Note:	*p<0.1; **p<0.05; ***p<0.01

From the estimation of the gravity equation controlling for all pair of countries who are in the Eurozone we see that the estimated coefficient on the EuroDummy is statistically significant at 5 % significance level and has also a positive sign. This shows that a pair countries who is part of Eurozone has a trade flow of 0.374 % higher than a pair of countries who is not part of the Eurozone. Hence the countries in the Eurozone have 0.374 % more trade when compared to countries not in the Eurozone. Also the other estimates are very close to the previous results.

We find that R2 in this model is almost the same with R2 with the previous model. This shows that the EuroDummy that we controlled for, does not explain much of the variation in the trade flow.

1.9. Do you think your estimate of the coefficient β_4 reflects the causal impact of the Euro on trade? Why or why not? Can you think of reasons why the error term u_ij might be correlated with the Euro dummy?

I do not think that the estimated coeffcient on beta_4 reflects the impact of the Euro on trade. The reason is due to simultaneity in the equation. We can raise the question: do these countries trade more because of the common currency or they joined the common currency because they trade more with each other. These countries have historical ties with each other, and that is something that we cannot control for, and some of them had bilateral trade agreements prior of joining the Eurozone. Hence simultaneity is a serious concern which causes endogeneity.

The gravity equation can only capture correlations and not causality. Hence the estimated coefficient must be biased and inconsistent due to simultaneity.

One concern of the above equation is that e.g. historical ties between EU countries or the membership in the same customs union affect both trade directly as well as the decision to enter the Eurozone. Since historical ties are for example difficult to measure, they will typically be part of the error term u_{ij} and OLS would give us a biased estimate of β_4. Therefore, instead of the above gravity equation, you are considering to compare trade before and after the introduction of the Euro using a diff-in-diff regression.

1.10. Collect the same data on trade flows as before but now for the year 1995. Match this data to country GDP in 1995 using ### the file "gdp 1995.RData"

Please note that due to some missing data for 1995 I used the data for 1996.

Exports (1996)

```
I_C <- as.character(c(gdp_1995$cty_code))
d1_96 <- NULL
for (i_c in I_C[1:5]) {
  for (i_k in I_C) {
    if (i_c == i_k) {
      next
    }
    else if (i_c != i_k) {
      data_i_c <- get.Comtrade(r = i_c, p = i_k, rg = "2", c = "TOTAL", ps = "1996")
      d1_96 <- rbind(d1_96, data_i_c)
    }
  }
  Sys.sleep(10)
```

```
}

Sys.sleep(3660)

I_C <- as.character(c(gdp_1995$cty_code))
d2_96 <- NULL
for (i_c in I_C[6:10]) {
  for (i_k in I_C) {
    if (i_c == i_k) {
      next
    }
    else if (i_c != i_k) {
      data_i_c <- get.Comtrade(r = i_c, p = i_k, rg = "2", c = "TOTAL", ps = "1996")
      d2_96 <- rbind(d2_96, data_i_c)
    }
  }
  Sys.sleep(10)
}

Sys.sleep(3660)

I_C <- as.character(c(gdp_1995$cty_code))
d3_96 <- NULL
for (i_c in I_C[11:15]) {
  for (i_k in I_C) {
    if (i_c == i_k) {
      next
    }
    else if (i_c != i_k) {
      data_i_c <- get.Comtrade(r = i_c, p = i_k, rg = "2", c = "TOTAL", ps = "1996")
      d3_96 <- rbind(d3_96, data_i_c)
    }
  }
  Sys.sleep(10)
}

Sys.sleep(3660)

I_C <- as.character(c(gdp_1995$cty_code))
d4_96 <- NULL
for (i_c in I_C[16:20]) {
  for (i_k in I_C) {
    if (i_c == i_k) {
      next
    }
    else if (i_c != i_k) {
      data_i_c <- get.Comtrade(r = i_c, p = i_k, rg = "2", c = "TOTAL", ps = "1996")
      d4_96<- rbind(d4_96, data_i_c)
    }
  }
  Sys.sleep(10)
}
```

Imports (1996)

```
I_C <- as.character(c(gdp_1995$cty_code))
d5_96 <- NULL
for (i_c in I_C[1:5]) {
  for (i_k in I_C) {
    if (i_c == i_k) {
      next
    }
    else if (i_c != i_k) {
      data_i_c <- get.Comtrade(r = i_c, p = i_k, rg = "1", c = "TOTAL", ps = "1996")
      d5_96 <- rbind(d5_96, data_i_c)
    }
  }
  Sys.sleep(10)
}

Sys.sleep(3660)

I_C <- as.character(c(gdp_1995$cty_code))
d6_96 <- NULL
for (i_c in I_C[6:10]) {
  for (i_k in I_C) {
    if (i_c == i_k) {
      next
    }
    else if (i_c != i_k) {
      data_i_c <- get.Comtrade(r = i_c, p = i_k, rg = "1", c = "TOTAL", ps = "1996")
      d6_96 <- rbind(d6_96, data_i_c)
    }
  }
  Sys.sleep(10)
}

Sys.sleep(3660)

I_C <- as.character(c(gdp_1995$cty_code))
d7_96 <- NULL
for (i_c in I_C[11:15]) {
  for (i_k in I_C) {
    if (i_c == i_k) {
      next
    }
    else if (i_c != i_k) {
      data_i_c <- get.Comtrade(r = i_c, p = i_k, rg = "1", c = "TOTAL", ps = "1996")
      d7_96 <- rbind(d7_96, data_i_c)
    }
  }
  Sys.sleep(10)
}

Sys.sleep(3660)

I_C <- as.character(c(gdp_1995$cty_code))
d8_96 <- NULL
```

```
for (i_c in I_C[16:20]) {
  for (i_k in I_C) {
    if (i_c == i_k) {
      next
    }
    else if (i_c != i_k) {
      data_i_c <- get.Comtrade(r = i_c, p = i_k, rg = "1", c = "TOTAL", ps = "1996")
      d8_96 <- rbind(d8_96, data_i_c)
    }
  }
  Sys.sleep(10)
}
```

Entire dataset for year 1996.

```
data_countries_1996 <- rbind(d1_96, d2_96, d3_96, d4_96, d5_96, d6_96, d7_96, d8_96)
data_countries_1996 <- data.table(data_countries_1996)

keycols <- c("cty_code_o","cty_code_d")
setnames(data_countries_1996, c("rtCode","ptCode"), keycols)
setnames(gdp_1995, "cty_code", "cty_code_o")
```

Convert column classes to numeric.

```
data_countries_1996$cty_code_o <- as.numeric(as.character(data_countries_1996$cty_code_o))
data_countries_1996$cty_code_d <- as.numeric(as.character(data_countries_1996$cty_code_d))

gdp_1995 <- data.table(gdp_1995)
data_countries_1996 <- data.table(data_countries_1996)
```

Please note that all the matchings was done using the GDP 1995

using the data on year 1996 for TradeFlow.

Match for gdp_1995_j

Match for gdp_1995_i

```
data_countries_1996$gdp_1995_i <- NA
I <- as.numeric(c(gdp_1995$cty_code_o))
K <- as.numeric(c(data_countries_1996$cty_code_o))

for (i in seq_along(I)) {
  for (j in seq_along(K)) {
    if (gdp_1995$cty_code_o[i] == data_countries_1996$cty_code_o[j]) {
      data_countries_1996$gdp_1995_i[j] <- gdp_1995$gdp_1995[i]
    }
    else if (gdp_1995$cty_code_o[i] != data_countries_1996$cty_code_o[j]) {
      next
    }
  }
}
```

```
data_countries_1996$gdp_1995_j <- NA
I <- as.numeric(c(gdp_1995$cty_code_o))
K <- as.numeric(c(data_countries_1996$cty_code_o))

for (i in seq_along(I)) {
  for (j in seq_along(K)) {
    if (data_countries_1996$cty_code_d[j] == gdp_1995$cty_code_o[i]) {
      data_countries_1996$gdp_1995_j[j] <- gdp_1995$gdp_1995[i]
    }
    else if (data_countries_1996$cty_code_d[j] != gdp_1995$cty_code_o[i]) {
      next
    }
  }
}
```

```
setnames(data_countries_1996, "TradeValue", "TradeValue_1996")
data_countries_1996 <- data_countries_1996[ , .(rgDesc, cty_code_o, cty_code_d, gdp_1995_i,
                                                gdp_1995_j, TradeValue_1996)]

data_countries_1996$TradeValue_1996 <- as.numeric(as.character(data_countries_1996$TradeValue_1996))
```

1.11. Add the 1995 **data to the previous one for** 2014**. You should end up with** 3 **new columns:**

(1) Trade flows for 1995, (2) GDP of country i in 1995 and (3) GDP of country j in 1995.

The following loop will merge based on: TradeFlow, cty_code_o, and cty_code_d.

```
M <- as.numeric(c(data_countries_2014$cty_code_o))
N <- as.numeric(c(data_countries_1996$cty_code_o))


for (i in seq_along(M)) {
  for (j in seq_along(N)) {
    if (data_countries_2014$cty_code_o[i] == data_countries_1996$cty_code_o[j] &&
        data_countries_2014$cty_code_d[i] == data_countries_1996$cty_code_d[j] &&
        data_countries_2014$TradeFlow[i] == data_countries_1996$TradeFlow_1996[j]) {

      data_countries_2014$TradeValue_1996[i] <- data_countries_1996$TradeValue_1996[j]
      data_countries_2014$gdp_1995_i[i] <- data_countries_1996$gdp_1995_i[j]
      data_countries_2014$gdp_1995_j[i] <- data_countries_1996$gdp_1995_j[j]

    }
  }
}
```

```
## Create a new dataset to use it for the DID modeling.
## Note that the first difference between the indicator function for euro in 2014
## minus the indicator function for euro 1995 is equal to the indicator function
## for euro in 2014.
```

```
diff_log_trade_value <- log(data_countries_2014$TradeValue) - log(data_countries_2014$TradeValue_1996)
diff_log_gdp_i <- log(data_countries_2014$gdp_2014_i) - log(data_countries_2014$gdp_1995_i)
diff_log_gdp_j <- log(data_countries_2014$gdp_2014_j) - log(data_countries_2014$gdp_1995_j)
diff_euro <- data_countries_2014$EuroDummy
```

The following dataset includes all the first differences.

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     .N
```

```
diff_data <-cbind(diff_log_trade_value, diff_log_gdp_i, diff_log_gdp_j, diff_euro )
diff_data <- data.table(diff_data)
```

1.12 In order to derive the diff-in-diff specification, first write down equation (2)

for both years, explicitly keeping track of the respective period.

$$ln(x_{ij}^{(14)}) = \beta_0 + \beta_1 \ln(y_i^{(14)}) + \beta_2 \ln(y_j^{(14)}) + \beta_3 \ln(dist_{ij}) \beta_4 I\{Euro^{(14)}\} + u_{ij^{(14)}}$$

$$ln(x_{ij}^{(95)}) = \beta_0 + \beta_1 \ln(y_i^{95}) + \beta_2 \ln(y_j^{(95)}) + \beta_3 \ln(dist_{ij}) \beta_4 I\{Euro^{(95)}\} + u_{ij^{(95)}}$$

The superscript numbers in brackets represent the respective years, i.e. 1995 and 2014. Notice that the error term is also allowed to differ by year, i.e. also unobservables that determine trade flows might change over time.Subtracting the second from the first line gives us a desired diff-in-diff version of the gravity equation:

$$\ln(x_{ij}^{(14)}) - ln(x_{ij}^{(95)}) = \beta_1 \big[\ln(y_i^{(14)}) - \ln(y_i^{95})\big] + \beta_2 \big[\ln(y_j^{(14)}) - \ln(y_j^{(95)})\big] + \beta_4 \big[I\{Euro^{(14)}\} - I\{Euro^{(95)}\} + \big] \big[u_{ij}^{(14)} - u_{ij}^{(95)}\big]$$

which can be written more compactly as

$$\Delta \ln(x_{ij}) = \beta_1 \Delta \ln(y_i) + \beta_2 \Delta \ln(y_j) + \beta_4 \Delta I\{Euro\} + \Delta u_{ij}$$

Estimate this equation again with the lm() function. What do you find? How do your results differ from those found in (8)?

First difference model.

```
first_diff_model <- lm(diff_log_trade_value ~ 0 + diff_log_gdp_i + diff_log_gdp_j + diff_euro,
                       diff_data)

stargazer(first_diff_model, type = "latex",
          title            = "Estimation of Gravity Equation with First Differencing",
          covariate.labels = c("$d (ln(y_i))$", "$d (ln(y_j))$", "EuroDummy"),
          dep.var.caption  = "Trade Value",
          dep.var.labels   = "$$", header = FALSE)
```

We find that the estimate on the Euro dummy variable is statistically insigificant when compared to the previous estimated equation. Hence using first differencing we were able to remove any fixed factors that vary across time. While the estimates on the GDP's have decreased in magnitude when compared to the previous estimated coefficients.

Table 3: Estimation of Gravity Equation with First Differencing

	Trade Value
$d(ln(y_i))$	0.695^{***}
	(0.033)
$d(ln(y_j))$	0.699^{***}
	(0.033)
EuroDummy	-0.097
	(0.117)
Observations	760
R^2	0.828
Adjusted R^2	0.827
Residual Std. Error	0.735 (df = 757)
F Statistic	$1{,}215.650^{***}$ (df = 3; 757)
Note:	*p<0.1; **p<0.05; ***p<0.01

1.13 Suppose the error term u_{ij} depends partially on factors that are time-invariant ν_{ij}, e.g. the language spoken in a country, and partially on factors that might vary over time

$\nu(t)$, e.g. tariffs and shipping costs:

$$u_{ij} = \nu_{ij} + \nu_{ij}^{(t)}$$

Suppose you are worried that sharing a common language affects both trade flows directly as well the decision to enter the Eurozone and hence $Corr(\nu_{ij}, IEuro) \neq 0$. Would this be a problem in regression (2)? And in the diff-in-diff specification?

If $Cov(\nu_{ij}, IEuro) \neq 0$ and $u_{ij} = \nu_{ij} + \nu_{ij}^{(t)}$ that implies that

$$Cov(u_{ij}, IEuro) = Cov(\nu_{ij} + \nu_{ij}^{(t)}, I\{Euro\}) = \underbrace{Cov(\nu_{ij} \times I\{Euro\})}_{\neq 0} + Cov(\nu_{ij}^{(t)}, I\{Euro\}) \neq 0$$

Hence as $Cov(v_{ij}, IEuro) \neq 0$ that implies that $Cov(u_{ij}, IEuro) \neq 0$. This results in a serious endogeneity problem for regression (2) which will result in biased and inconsistent estimates. This is an omitted variable bias problem as we fail to control for time invariant and time variant factors which sometimes are unobserved to the econometrician.

However in the DID model specification we have that

$$\begin{aligned} Cov(\Delta u_{ij}, \Delta I\{Euro\}) &= Cov\big(u_{ij}^{(14)} - u_{ij}^{(95)}, \Delta I\{Euro\}\big) \\ &= Cov\big(\nu_{ij}^{(14)} + \nu_{ij}^{(t)} - (\nu_{ij}^{(95)} + \nu_{ij}^{(t)}), \Delta I\{Euro\}\big) \\ &= Cov\big(\nu_{ij}^{(14)} - \nu_{ij}^{(95)}, \Delta I\{Euro\}\big) \\ &= \underbrace{Cov\big(\nu_{ij}^{(14)} \times \Delta I\{Euro\}\big)}_{\neq 0} + \underbrace{Cov\big(\nu_{ij}^{(95)} \times \Delta I\{Euro\}\big)}_{\neq 0} \end{aligned}$$

or to simplify the analysis we can set $\nu_{ij}^{(14)} - \nu_{ij}^{(95)} = z_{ij}$, then

$$= Cov\big(z_{ij}, \Delta I\{Euro\}\big) \neq 0$$

Hence the first difference model removed the time fixed effects however the time invariant effects are still in the model. From the estimation results we can see that the estimated coefficient on EuroDummy is statistically insiginificant.

Would you now interpret the estimate for β_4 as the causal impact of the Euro on trade. Why or why not? Can you think of a case in which Δu_{ij} might still be correlated with $\Delta I\{Euro\}$?

After first differencing I am willing to interpret the estimate for β_4 as association but no causality. Even if we applied first differencing to remove fixed time effects there are still other time invariant factors that should be considered. As I show above Δu_{ij} might still be correlated with $\Delta I\{Euro\}$, and this will happen due to the time invariant effects which change across pair of countries but do not change with time. A case I would think could be historical ties that countries may have had prior of joining the Eurozone. The European countries in the dataset which are part of the Eurozone long before the Euro was introduced were part of the European Economic Area which is a free movement of persons, goods, services and capital within the European Single Market. So I believe that Δu_{ij} might still be correlated with $\Delta I\{Euro\}$.