

Assignment 2

Bilal Shakir

2018-01-18

Assignment 1

Introduction

This assignment is based on “The design of international trade agreements” dataset that was introduced in Dür, Andreas, Leonardo Baccini, and Manfred Elsig. (2014) “The design of international trade agreements: Introducing a new dataset.” *The Review of International Organizations* 9(3): 353-375.

Before proceeding with the actual tasks set forth in the assignment 2 I set up my data and variable so that they are ready for analysis. In the following code chunk, I load the necessary packages used in this assignment and rewrite the dataset in a format that is analysed more efficiently by R (.RDS) using the <> package. The faster computation of .RDS format (in conjunction with more efficient code) allows allows me to analyze the entire dataset rather than forcing me to run my analysis on subsets of the data.

```
## clearing out the global environment is a good idea before starting out
```

```
remove(list=ls())
```

```
## loading in the required datasets
```

```
pacman::p_load(tidyr, tidyverse,
               ggplot2, ggthemes,
               dplyr, RColorBrewer,
               gridExtra,
               texreg, readr,
               foreign, readstata13,
               interplot, stargazer,
               Zelig, broom,
               car, purrr,
               MASS, arm,
               pander, knitr,
               psych, cowplot,
               margins, scales,
               pastecs, plm, survival,
               ggplot2, ggrepel, pscl,
               GGally, foreign, RTextTools,
               feather,
               lfe, dplyr, corrplot, haven)
```

```
## renaming dataset for ease
```

```
gravity_data <- read_dta("gravitydata_assign3_2018.dta")
```

```
write_rds(gravity_data, "gravity_rds.rds")
```

```
gravity_data <- read_rds("gravity_rds.rds")
```

```
# gravity_small <- subset(gravity_small, cow1 > 300 | cow2 >200 )
```

Transforming the required variables for the Gravity model

In the following code chunk I transform the variables that are typically transformed in the gravity model. This transformation is necessary for several reasons, chiefly those relating to having the variables in a format so that they are amenable to a linear model (such as OLS). For the sake of brevity, I do not show plots that demonstrate and justify the transformation of these variables. It is assumed instead that the transformation of these variables is standard in the gravity model and as done by the paper which provided the substantive and methodological inspiration behind this assignment “Baier, S. L., and Bergstrand, J. H. (2007). “Do free trade agreements actually increase members’ international trade?” *Journal of international Economics*, 71(1), 72-95.”

After log-transforming the variables I store their descriptive statistics in the following code-chunk.

```
## first let's take a look at the broad dataset
```

```
glimpse(gravity_data)
```

```
## Observations: 1,955,288
```

```
## Variables: 65
```

```
## $ cow1      <dbl> 310, 290, 140, 390, 450, 840, 290, 145, 678, 9...
## $ cow2      <dbl> 57, 100, 660, 315, 225, 232, 40, 552, 710, 680...
## $ year      <dbl> 1948, 1948, 1948, 1948, 1948, 1948, 1948, 1948...
## $ M         <dbl> NA, 0.00, 0.00, 10.70, NA, NA, 0.11, NA, NA, 0...
## $ X         <dbl> 0.0, 0.0, 0.0, 9.0, NA, NA, 0.0, 0.0, NA, 0.0,...
## $ expab     <dbl> NA, 0.0, 0.0, 9.0, 0.0, 0.0, 0.0, NA, 0.0, NA,...
## $ eabo      <dbl> NA, 1, 0, 0, 1, 6, 1, NA, 1, NA, NA, NA, 2, 2,...
## $ impab     <dbl> NA, 0.0, 0.0, 11.0, 0.0, 0.0, 0.1, NA, 0.0, NA...
## $ iabo      <dbl> NA, 2, 2, 0, 2, 6, 0, NA, 2, NA, NA, NA, 2, 2,...
## $ democ_A   <dbl> 0, 0, 6, 10, 0, 4, 0, 1, 1, 1, -88, 10, 0, 10,...
## $ autoc_A   <dbl> 7, 7, 1, 0, 6, 2, 7, 6, 7, 4, -88, 0, 6, 0, 0,...
## $ polity_A  <dbl> -7, -7, 5, 10, -6, 2, -7, -5, -6, -3, -88, 10,...
## $ polity2_A <dbl> -7, -7, 5, 10, -6, 2, -7, -5, -6, -3, -3, 10, ...
## $ democ_B   <dbl> NA, 1, 4, 0, 10, NA, 4, NA, 1, NA, NA, NA, 3, ...
## $ autoc_B   <dbl> NA, 6, 2, 7, 0, NA, 1, NA, 6, NA, NA, NA, 3, -...
## $ polity_B  <dbl> NA, -5, 2, -7, 10, NA, 3, NA, -5, NA, NA, NA, ...
## $ polity2_B <dbl> NA, -5, 2, -7, 10, NA, 3, NA, -5, NA, NA, NA, ...
## $ tomz_gdp_67_B <dbl> NA, 3510054289, 589131071, 5064527775, 5978988...
## $ dtomz_gdp_67_B <dbl> NA, 0.087141842, NA, -0.128557295, -0.01610577...
## $ tomz_gdp_67_A <dbl> 1871028916, 7862180932, 14255292695, 511308623...
## $ dtomz_gdp_67_A <dbl> 0.251276821, 0.320388347, -0.011455184, 0.0085...
## $ cpi_A     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ gdp_wb_A  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ cpi_B     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ gdp_wb_B  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ uscpi     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ uscpi67   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ newgdp_67_A <dbl> 1871028916, 7862180932, 14255292695, 511308623...
## $ newgdp_67_B <dbl> NA, 3510054289, 589131071, 5064527775, 5978988...
## $ pop_pwt_A <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ rgdpch_A  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ pop_pwt_B <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ rgdpch_B  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ pop_madd_A <dbl> 9158, 23980, 49590, 4190, NA, 19918, 23980, 29...
## $ pop_madd_B <dbl> NA, 11332, NA, 12339, 4582, NA, 5268, NA, 5410...
## $ pop_A     <dbl> 9158, 23980, 49590, 4190, NA, 19918, 23980, 29...
## $ pcgdp_A   <dbl> 204.3054, 327.8641, 287.4630, 1220.3070, NA, 2...
```

```
## $ pop_B          <dbl> NA, 11332, NA, 12339, 4582, NA, 5268, NA, 5410...
## $ pcgdp_B        <dbl> NA, 309.7471, NA, 410.4488, 1304.8862, NA, 629...
## $ wtojoinA        <dbl> 1995, 1995, 1995, 1995, NA, 1995, 1995, 1995, ...
## $ wtojoinB        <dbl> 1995, 1995, NA, NA, 1995, NA, 1995, 1995, 2001...
## $ gattjoinA       <dbl> 1973, 1967, 1948, 1950, NA, 1979, 1967, 1990, ...
## $ gattjoinB       <dbl> 1993, 1981, NA, 1948, 1966, NA, 1948, 1948, NA...
## $ onegatt         <dbl> 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1...
## $ gattwto         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ lat_1           <dbl> 47.500000, 52.250000, -15.780000, 55.666672, 6...
## $ lon_1           <dbl> 19.049999, 21.000000, -47.910000, 12.583330, -...
## $ lat_2           <dbl> 13.1999998, 4.6333332, 33.8666611, 50.0999985,...
## $ lon_2           <dbl> -61.233330, -75.066673, 35.500000, 14.433330, ...
## $ distcap2        <dbl> 8203.6795, 10011.4947, 10389.0310, 631.2497, 4...
## $ contig          <dbl> 0, 0, 0, NA, 0, 0, 0, 0, NA, NA, 0, 0, 0, 0, 0...
## $ dist            <dbl> 8211.371, 10020.882, 10772.672, NA, 4843.690, ...
## $ distcap         <dbl> 8211.371, 10020.882, 10398.772, NA, 4843.690, ...
## $ distw           <dbl> 8227.310, 9790.112, 10299.355, NA, 4850.261, 1...
## $ distwces        <dbl> 8226.332, 9781.810, 10241.749, NA, 4848.342, 1...
## $ conttype        <dbl+lbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ cowcontig       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ depth_rasch_new <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ flex_rasch_new  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ enf_rasch_new   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ id              <dbl> 12553, 12113, 7637, 19527, 22744, 41990, 12090...
## $ pta             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ name            <chr> "", "", "", "", "", "", "", "", "", "", "", "", ""...
## $ reg_europe      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ pta1            <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
```

There are a total of 65 variables in the dataset. This makes the dataset too large to be described fully.

Consequently, let's describe the variables of interest

```
#stat.desc(gravity_data$M)
```

we don't even need to plot the imports variable to see that the imports data needs to be log transformed. This is standard for the gravity model.

```
gravity_data$ln_M <- log(gravity_data$M+1)
```

Descriptive stats for Imports

```
#desc_ln_M <- stat.desc(gravity_data$ln_M); desc_ln_M
```

transforming GDP Country A

```
gravity_data$ln_newgdp_67_A <- log(gravity_data$newgdp_67_A)
```

transforming GDP B

```
gravity_data$ln_newgdp_67_B <- log(gravity_data$newgdp_67_B)
```

transforming distance

```
gravity_data$ln_dist <- log(gravity_data$dist)

# desc_ln_M
# desc_gdp_A
# desc_gdp_B
# desc_ln_dist

## transforming the rest of the variables

#desc_depth <- stat.desc(gravity_data$depth_rasch_new); desc_depth
#desc_polity2_A <- stat.desc(gravity_data$polity2_A); desc_polity2_A
#desc_polity2_B <- stat.desc(gravity_data$polity2_B); desc_polity2_B
#desc_gattwo <- stat.desc(gravity_data$gattwo); desc_gattwo
#desc_contig <- stat.desc(gravity_data$contig); desc_contig
```

Task 1.1

Check the correlation among the main explanatory variables and decide an appropriate model specification.

Solution 1.1

One of the first main explanatory variable in the DESTA dataset is *depth_rasch_new*. This is a variable that is created via “latent trait analysis on a total of 48 variables that theoretically are related to the depth of an agreement (these variables pertain to such aspects as services liberalization, trade-related investment measures, intellectual property rights and standards) to arrive at a measure of depth.” Moreover, note that Latent trait analysis can usefully be conceived as a form of factor analysis for binary data (Bartholomew et al. 2011).

More concretely, Dur et al. (2014) then apply the Rasch model that assumes that all items capture one underlying latent dimension, but with different discriminatory power. Doing so allows them to deal with highly correlated data and to account for the fact that not all items are of equal importance in establishing the extent of countries’ commitments. Ultimately, this results in the creation of the variable *depth_rasch_new* which a crucial explanatory variable for our analysis.

Similar to the process behind the creation of *depth_rasch_new*, two other variables were also created by the authors. These include *flex_rasch_new* and *enf_rasch_new* that measure the flexibility and enforcement of PTAs respectively. Note that these are continuous variables created using latent trait analysis. So in a certain sense the creation of these variables can be thought of as applying a method onto binary variables (PTAs, for example) and converting them into continuous measures.

In summary, depth (or *depth_rasch_new*) is basically the amount of things we are going to say that a preferential trade agreement is going to affect. The more deep a PTA between two countries the more domestic policy of these countries is affected by it. Whereas, flexibility (or *flex_rasch_new*) refers to how much room you have maneuver within the commitments of a trade agreement. Finally, enforcement mechanisms (or *enf_rasch_new*) are the enforcement mechanisms that are in place to ensure that things follow up or the institutions or clauses that are in place to make sure that countries follow up on their commitments of a trade agreement. This background will be useful in analyzing the correlation plots in the analysis that follows.

Correlation plot for main explanatory variables

Figure 1 below is a correlation plot that shows that there is a high degree of correlation between all three indexes aforementioned. This implies that the three indexes are endogeneous to each other. As such, it is important not to include all three measures as covariates in a single linear model as that would be a violation of a pooled OLS model. Instead, for the course of this assignment I will run three different linear models, each with one of the three main explanatory variables to estimate trade levels through a linear model. However,

for the sake of brevity, I will only discuss in detail the results that I obtain using *depth_rasch_new* as the main explanatory variable.

Note that my correlation plot between the three main explanatory variables shows that there is a high degree of correlation between depth, flexibility and enforcement of trade agreements. More specifically, these high correlations between these three variables vary between (0.9 and 0.77). However this high degree of correlation is not too surprising between these variables. This is because the deeper the agreement, the more likely it is that there would be some sort of flexibility required, in the form of escape laws during the negotiating of trade agreements between two countries. This is mentioned by Johns when he underlines that “Deep tariff bindings will be accompanied by flexibility provisions, while more rigid agreements will have shallower trade bindings” (2014, 472).

Similarly, the enforcement mechanism that follows a trade agreement is likely to vary according to the trade agreements depth and flexibility. This is the reason for running models with each of these explanatory variables separately rather than including them all at once in one model.

Correlation plot for baseline variables (and some additional controls)

Figure 2 is a correlation plot that shows the correlation between all the main variables in our baseline gravity model. First, let’s discuss the main variables that provide evidence of a positive correlation. We can observe that there is some correlation between the polity of a country and gdp (in this case, our data seems to indicate a positive relationship between democracies and gdp, in other words democracies will tend to have a higher economic growth). There is also a positive correlation between gattwto and polity. Our data provides some evidence that countries with higher polity scores (or democracies) will be likely to both be members of GATT/WTO.

More crucially, there is some positive correlation between pta and the explanatory variables of depth, flexibility and enforcement (BETWEEN 0.9 to 0.77). Due to this high correlation I do not include PTA in the least squares models that I estimate in the following questions. This is contrary to Berg and Bergstrand (2007), however, as pta’s are endogenous to our main explanatory variables, pta’s inclusion should be avoided.

The only variables that provide evidence of a negative correlation based on Figure 2 that are notable include distance and contig. However, this seems intuitive as it makes sense that countries that have a larger distance between their capital cities will not share borders. But as the coefficient of this correlation (-0.4) is much lower than that for PTAs and the main explanatory variables, I retain both distance and contig in my main models. Here, I make this decision by choosing a correlation of 0.5 for positive and -0.5 as the cut-off for determining whether a variable should be included in the model or not. In other words, if the correlation coefficient is lower than 0.5 or higher than -0.5 they are retained in the least squares models.

The model should therefore include the following variables, generally included in gravity models: import, GDP, distance, contiguity, regime, and GATT/WTO. The measures are the variables included in the data set provided for this assignment:

LHS: $-\ln_M$ (outcome)

RHS: $-\text{depth_rasch_new} - \ln_newgdp_67_A - \ln_newgdp_67_B - \text{polity2_A} - \text{polity2_B} - \text{gattwto} - \ln_dist - \text{contig}$

```
## the main explanatory variable
```

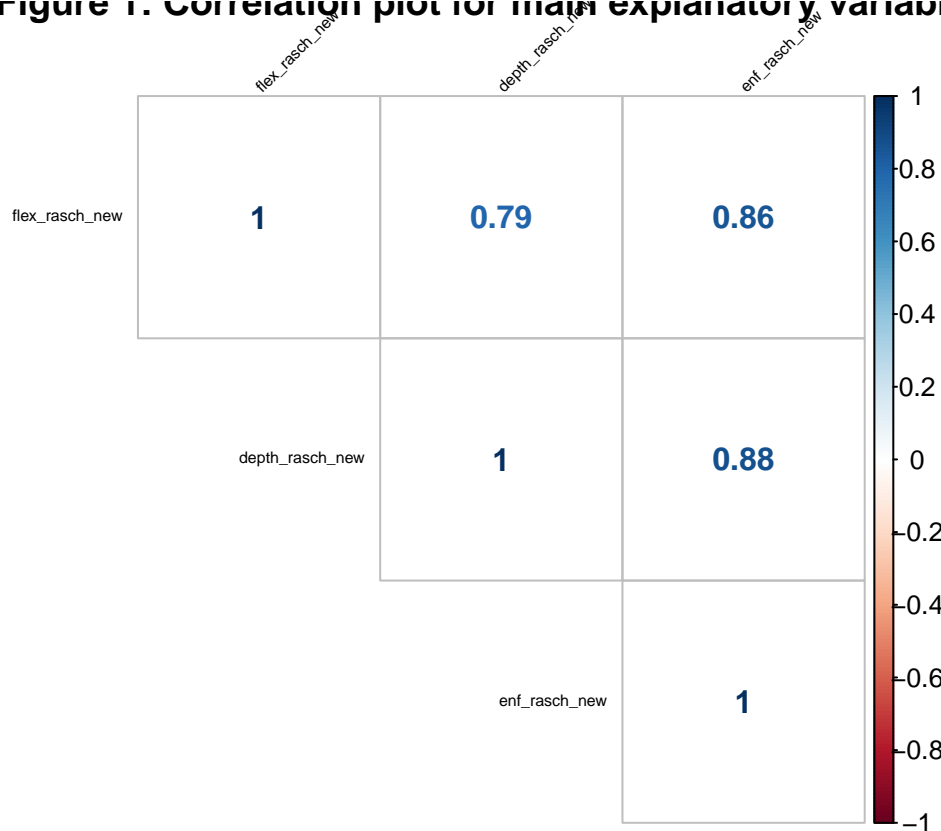
```
#gravity_data$depth_rasch_new
#gravity_data$flex_rasch_new
#gravity_data$enf_rasch_new
```

```
### correlation plot for main explanatory variables and the indexes
```

```
gravity_data %>%
  dplyr::select(depth_rasch_new, flex_rasch_new, enf_rasch_new) %>%
  cor(use = "complete.obs") %>%
```

```
corrplot(type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45,
         tl.cex = 0.5, method = "number",
         title = "Figure 1: Correlation plot for main explanatory variables")
```

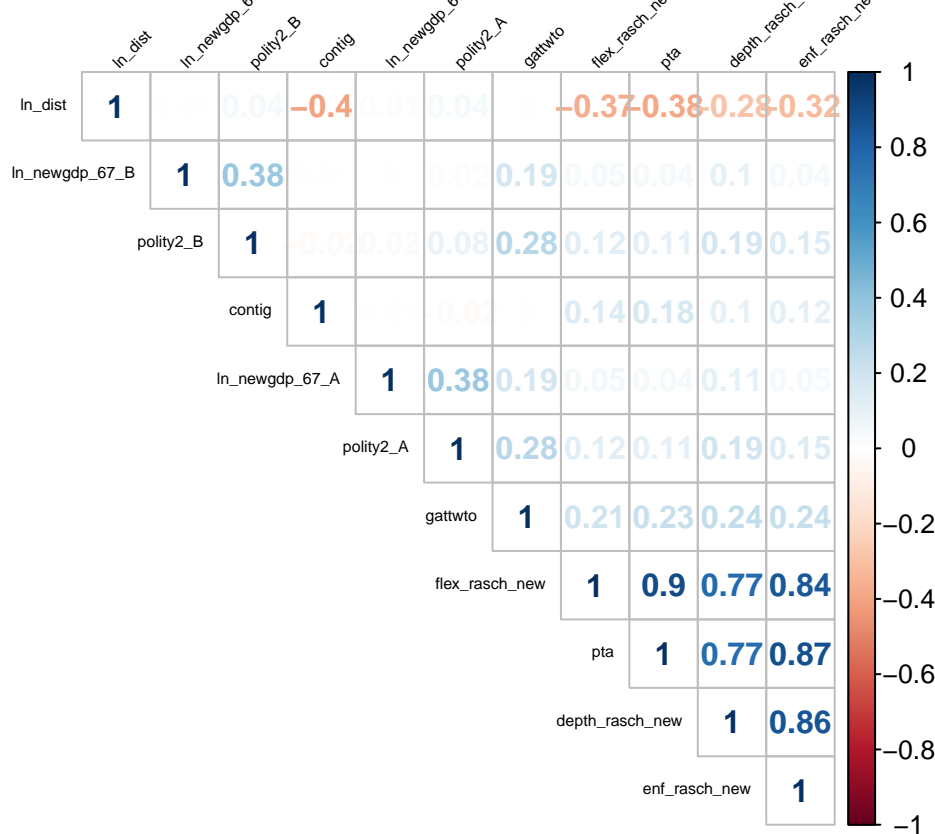
Figure 1. Correlation plot for main explanatory variables



```
## correlation plot between all the variables in the baseline model (and some additional ones)

gravity_data %>%
  dplyr::select(depth_rasch_new, flex_rasch_new,
                enf_rasch_new, ln_newgdp_67_A,
                ln_newgdp_67_B, polity2_A,
                polity2_B, gattwto,
                ln_dist, contig, pta) %>%
  cor(use = "complete.obs") %>% corrplot(type = "upper",
                                         order = "hclust",
                                         tl.col = "black",
                                         tl.srt = 45,
                                         tl.cex = 0.5,
                                         method = "number",
                                         title = "Figure 2: Correlation plot for baseline variables")
```

Figure 2. Correlation plot for baseline variables



```
# include pta, pta1
```

Task 1.2

2. Run a simple OLS with the main gravity model variables. Explain the results.

Solution 1.2

In the following code chunk I run a simple OLS regression with the main gravity model variables. Note that the decision behind the inclusion (and omission) of certain variables is discussed in greater detail in the analysis section that follows the code chunk below. Note that the variables have been transformed where applicable (some of rationale behind this is explained in solution 1.1 earlier, though in short, I transform variables based on standard practice for a gravity model based on Baier and Bergstrand and Baccini et al 2012)

```
## let's describe the variables of interest
```

```
stat.desc(gravity_data$M)
```

```
##              x
## nbr.val      1.289722e+06
## nbr.null     6.016250e+05
## nbr.na       6.655660e+05
## min          0.000000e+00
## max          3.563190e+05
## range        3.563190e+05
## sum          1.731672e+08
```

```
## median      7.977790e-03
## mean        1.342671e+02
## SE.mean     1.873964e+00
## CI.mean.0.95 3.672906e+00
## var         4.529170e+06
## std.dev     2.128185e+03
## coef.var    1.585038e+01
```

```
## Descriptive stats for Imports
```

```
stat.desc(gravity_data$M)
```

```
##              x
## nbr.val      1.289722e+06
## nbr.null     6.016250e+05
## nbr.na       6.655660e+05
## min          0.000000e+00
## max          3.563190e+05
## range        3.563190e+05
## sum          1.731672e+08
## median       7.977790e-03
## mean         1.342671e+02
## SE.mean      1.873964e+00
## CI.mean.0.95 3.672906e+00
## var          4.529170e+06
## std.dev      2.128185e+03
## coef.var     1.585038e+01
```

```
## transforming GDP Country A
```

```
gravity_data$ln_gdp_wb_A <- log(gravity_data$newgdp_67_A)
```

```
# descriptive
```

```
stat.desc(gravity_data$newgdp_67_A)
```

```
##              x
## nbr.val      1.896721e+06
## nbr.null     0.000000e+00
## nbr.na       5.856700e+04
## min          1.309342e+06
## max          2.253637e+12
## range        2.253636e+12
## sum          5.300699e+16
## median       1.823497e+09
## mean         2.794665e+10
## SE.mean      9.122055e+07
## CI.mean.0.95 1.787891e+08
## var          1.578297e+22
## std.dev      1.256303e+11
## coef.var     4.495362e+00
```

```
# transforming GDP B
```

```
gravity_data$ln_gdp_wb_B <- log(gravity_data$newgdp_67_A)
```

```
## transforming distance
```



```

gravity_data$ln_dist <- log(gravity_data$dist)

## Model 1 - with depth

model_1 <- lm(ln_M ~ depth_rasch_new + ln_newgdp_67_A +
              ln_newgdp_67_B + polity2_A + polity2_B +
              gattwto + ln_dist + contig,
              data = gravity_data)

## Model 2 - with flexibility

model_2 <- lm(ln_M ~ flex_rasch_new + ln_newgdp_67_A +
              ln_newgdp_67_B + polity2_A + polity2_B +
              gattwto + ln_dist + contig,
              data = gravity_data)

## Model 3 - with enforcement

model_3 <- lm(ln_M ~ enf_rasch_new + ln_newgdp_67_A +
              ln_newgdp_67_B + polity2_A + polity2_B +
              gattwto + ln_dist + contig,
              data = gravity_data)

### All three models together

table_1 <- screenreg(list(model_1, model_2, model_3),
                      stars = 0.05,
                      custom.note = "Standard errors in parentheses. *p < 0.05",
                      caption = "Table 1: Pooled OLS with the main gravity model variables"); table_1

```

```

##
## =====
##              Model 1      Model 2      Model 3
## -----
## (Intercept)    -16.10 *    -16.47 *    -16.48 *
##                (0.03)      (0.03)      (0.03)
## depth_rasch_new  0.26 *
##                (0.00)
## ln_newgdp_67_A   0.48 *      0.48 *      0.48 *
##                (0.00)      (0.00)      (0.00)
## ln_newgdp_67_B   0.55 *      0.55 *      0.55 *
##                (0.00)      (0.00)      (0.00)
## polity2_A        0.02 *      0.02 *      0.02 *
##                (0.00)      (0.00)      (0.00)
## polity2_B        0.02 *      0.02 *      0.02 *
##                (0.00)      (0.00)      (0.00)
## gattwto          0.15 *      0.14 *      0.13 *
##                (0.00)      (0.00)      (0.00)
## ln_dist          -0.62 *     -0.60 *     -0.60 *
##                (0.00)      (0.00)      (0.00)
## contig           0.56 *      0.55 *      0.55 *
##                (0.01)      (0.01)      (0.01)
## flex_rasch_new           0.30 *

```

```
##                                (0.00)
## enf_rasch_new                                0.27 *
##                                (0.00)
## -----
## R^2                                0.63            0.63            0.64
## Adj. R^2                            0.63            0.63            0.64
## Num. obs.                        771688            771688            771688
## RMSE                              1.26            1.26            1.26
## =====
## Standard errors in parentheses. *p < 0.05
```

Model 1

Table 1 above shows the results from our three models. From Model 1 above, we can observe that all the main gravity model variables are statistically significant at the 95 percent significance level. Moreover, the direction of all the variables also falls in line with our expectations (For example, Country Year GDP is +ve and distance is -ve).

However, as Baier and Bergstrand underline, there is a potential issue of endogeneity within our model which is not captured by a pooled OLS model. There is reason to suspect that trade agreements are endogenous to trade levels. For example, negotiations for trade agreements start much earlier before they actually come into effect. This implies that there are definitely anticipatory effects of actors of an upcoming trade agreement and it is possible that trade levels increase because of an expectation of a trade agreement rather than its actual implementation. Such a scenario would be a violation of the assumptions underlying an OLS model specification.

Crucially, this is the assumption that errors are uncorrelated with the dependent variables in an OLS modelling set-up. This assumption can be violated through three main things: 1) measurement error, 2) omitted variable bias and finally, 3) endogeneity. Here endogeneity can be understood as simultaneity or reverse causality wherein the two variables of interest are codetermined, with each affecting the other. So in that sense, our model has the problem of reverse causality which is overestimating the coefficient on the explanatory variables.

As underlined earlier, note that dyads with a PTA have a positive score of (higher than 0, i.e. depth is strictly positive for dyads with a PTA.) As such, unlike the advice of the Dur et al (2014) read for this class adding the PTA is not necessary because depth captures already the treatment units. Moreover, there was a high positive correlation between the PTA dummy and depth. This is why there is no need to include PTAs on the RHS of our models. By the standards of the gravity model, note that we are also getting a relatively low R squared and adjusted r-squared values. As endogeneity (such as serial correlation) is not really compared for under the OLS set up, we need a different modelling specification to get more reliable estimates.

Task 1.3

3. Add dyad and year (dummy) fixed-effects, modifying the model specification if necessary. Explain the results.

Solution 1.3

In the code chunk below I add dyad and year fixed effects to my models by creating new factor variables (as fixed effects). Note that I use the `<>` package in order to run the fixed effects linear models in this assignment for optimization as the `felm` package is very efficient than other alternatives (like `plm` or `lm`). As such, it allows me to estimate the coefficients for my covariates of interest (and outcome variable) on the entire data and not subsets of the data.

Note that *id* variable can be a fixed effect for the individual dyad. Whereas, the *cow* or correlates of war variable can be a fixed effect for country. Finally, the *year* variable is self explanatory and is a fixed effect for time or the year. Both the *id* and *year* variables are included in the linear models to create dyadic-year fixed

effects. Whereas I also report the country, year fixed effects by including *cow1* and *cow2* to capture country effects, and *year* to capture year fixed effects.

```
## adding new factor variables

## fixed effects for dyad
gravity_data$id <- as.factor(gravity_data$id)

## time country fixed country
gravity_data$cow1 <- as.factor(gravity_data$cow1)
gravity_data$cow2 <- as.factor(gravity_data$cow2)

## fixed effects for time (year)
gravity_data$year <- as.factor(gravity_data$year)

## dyad + year fixed effects models

model_4 <- felm (data = gravity_data, ln_M ~ depth_rasch_new +
                 ln_newgdp_67_A + ln_newgdp_67_B + polity2_A +
                 polity2_B + gattwto | id + year)

model_5 <- felm (data = gravity_data, ln_M ~ flex_rasch_new +
                 ln_newgdp_67_A + ln_newgdp_67_B + polity2_A +
                 polity2_B + gattwto | id + year)

model_6 <- felm (data = gravity_data, ln_M ~ enf_rasch_new +
                 ln_newgdp_67_A + ln_newgdp_67_B + polity2_A +
                 polity2_B + gattwto | id + year)

table_2 <- screenreg(list(model_4, model_5, model_6),
                      stars = 0.05,
                      custom.note = "Standard errors in parentheses. *p < 0.05",
                      caption = "Table 2: Dyad-year fixed effects"); table_2
```

```
##
## =====
##               Model 1      Model 2      Model 3
## -----
## depth_rasch_new      0.17 *
##                      (0.00)
## ln_newgdp_67_A       0.48 *      0.49 *      0.49 *
##                      (0.00)      (0.00)      (0.00)
## ln_newgdp_67_B       0.50 *      0.51 *      0.51 *
##                      (0.00)      (0.00)      (0.00)
## polity2_A            0.00 *      0.00 *      0.00 *
##                      (0.00)      (0.00)      (0.00)
## polity2_B           -0.00 *     -0.00 *     -0.00 *
##                      (0.00)      (0.00)      (0.00)
## gattwto              0.05 *      0.05 *      0.05 *
##                      (0.00)      (0.00)      (0.00)
## flex_rasch_new              0.16 *
##                          (0.00)
## enf_rasch_new              0.14 *
##                          (0.00)
## -----
```

```

## Num. obs.          828887      828887      828887
## R^2 (full model)    0.85        0.85        0.85
## R^2 (proj model)    0.09        0.09        0.09
## Adj. R^2 (full model) 0.84        0.84        0.84
## Adj. R^2 (proj model) 0.07        0.07        0.07
## =====
## Standard errors in parentheses. *p < 0.05
## Comparison of OLS vs Fixed Effects models

table_3 <- stargazer(model_1, model_4,
  header = FALSE, title = "Table 3: Comparison of Pooled OLS and Fixed Fixed Effects",
  keep = c("depth_rasch_new", "ln_newgdp_67_A", "ln_newgdp_67_B",
    "polity2_A", "polity2_B", "gattwto", "ln_dist", "contig"),
  dep.var.caption = "Dependent Variable: Log Imports (ln_M)",
  model.names = FALSE,
  column.labels = c("Pooled OLS", "+Dyad and Year FE"),
  dep.var.labels.include = FALSE,
  font.size = "scriptsize",
  model.numbers = FALSE, type = "text"); table_3

```

```

##
## Table 3: Comparison of Pooled OLS and Fixed Fixed Effects
## =====
##                               Dependent Variable: Log Imports (ln
##                               -----
##                               Pooled OLS          +Dyad and Year FE
##                               -----
## depth_rasch_new              0.264***           0.173***
##                               (0.003)            (0.002)
##
## ln_newgdp_67_A              0.476***           0.485***
##                               (0.001)            (0.003)
##
## ln_newgdp_67_B              0.549***           0.501***
##                               (0.001)            (0.002)
##
## polity2_A                    0.016***           0.004***
##                               (0.0002)           (0.0003)
##
## polity2_B                    0.018***           -0.002***
##                               (0.0002)           (0.0003)
##
## gattwto                      0.147***           0.055***
##                               (0.003)            (0.003)
##
## ln_dist                      -0.621***
##                               (0.002)
##
## contig                       0.557***
##                               (0.010)
## -----
## Observations                 771,688           828,887
## R2                           0.634            0.848

```

```

## Adjusted R2                                0.634                                0.844
## Residual Std. Error          1.259 (df = 771679)          0.823 (df = 804079)
## F Statistic          166,994.300*** (df = 8; 771679)
## =====
## Note:                                *p<0.1; **p<0.05; ***p<0.01

## [1] ""
## [2] "Table 3: Comparison of Pooled OLS and Fixed Fixed Effects"
## [3] "=====
## [4] "                                Dependent Variable: Log Imports (ln            "
## [5] "                                -----"
## [6] "                                Pooled OLS                                +Dyad and Year FE "
## [7] "-----"
## [8] "depth_rasch_new          0.264***          0.173***          "
## [9] "                        (0.003)          (0.002)          "
## [10] "
## [11] "ln_newgdp_67_A          0.476***          0.485***          "
## [12] "                        (0.001)          (0.003)          "
## [13] "
## [14] "ln_newgdp_67_B          0.549***          0.501***          "
## [15] "                        (0.001)          (0.002)          "
## [16] "
## [17] "polity2_A          0.016***          0.004***          "
## [18] "                        (0.0002)          (0.0003)          "
## [19] "
## [20] "polity2_B          0.018***          -0.002***          "
## [21] "                        (0.0002)          (0.0003)          "
## [22] "
## [23] "gattwto          0.147***          0.055***          "
## [24] "                        (0.003)          (0.003)          "
## [25] "
## [26] "ln_dist          -0.621***          "
## [27] "                        (0.002)          "
## [28] "
## [29] "contig          0.557***          "
## [30] "                        (0.010)          "
## [31] "
## [32] "-----"
## [33] "Observations          771,688          828,887          "
## [34] "R2          0.634          0.848          "
## [35] "Adjusted R2          0.634          0.844          "
## [36] "Residual Std. Error          1.259 (df = 771679)          0.823 (df = 804079)"
## [37] "F Statistic          166,994.300*** (df = 8; 771679)          "
## [38] "=====
## [39] "Note:                                *p<0.1; **p<0.05; ***p<0.01"

```

```
## Here I also report the Year, and country1 and country2 fixed effects
```

```

model_7 <- feLM(ln_M ~ depth_rasch_new + ln_newgdp_67_A +
                ln_newgdp_67_B + polity2_A + polity2_B +
                gattwto | cow1 + cow2 + year, data = gravity_data)

model_8 <- feLM(ln_M ~ flex_rasch_new + ln_newgdp_67_A +
                ln_newgdp_67_B + polity2_A + polity2_B +
                gattwto | cow1 + cow2 + year, data = gravity_data)

```

```
model_9 <- feIm(ln_M ~ enf_rasch_new + ln_newgdp_67_A +
               ln_newgdp_67_B + polity2_A + polity2_B +
               gattwto | cow1 + cow2 + year, data = gravity_data)
```

As per the class lectures, there are three main ways in which fixed effects may be implemented: 1) estimating with dummies, 2) within transformation and 3) first differencing. Table 2 shows results from a modelling specification that uses Fixed Effects by estimating c_i with dummies (*id* and *year*). I throw in these fixed effects because I know that it is silly to compare two different country dyads and two different years. Therefore, I capture everything specific to the dyad that is time-invariant by throwing in the *id* dummy. Whereas, the inclusion of the *year* dummy captures everything that is specific to that year. However there is still variation at the level of country years.

Moreover, as we had noted earlier that there might be omitted variable bias in models, such as what Baier underlines that often policy related to domestic regulations (barriers) is not observable to econometricians. This omitted variable is correlated to both trade and FTA (positively correlated) Baier and Bergstrand. Domestic regulations will reduce trade and increase the likelihood of trade. This is why we should fixed effects to account for this omitted variable bias.

We notice that the adjusted R-squared values increase from 0.634 in the Pooled OLS model to 0.844 in the dyad-year fixed effects model. Crucially, with regards to our main explanatory variable of interest depth, we observe that in line with our expectations discussed in the previous question, OLS was over estimating the coefficient of depth. We observe that with the inclusion of dyad and year fixed effects the coefficient on depth decreases from 0.264 to 0.173. Moreover, note that the control variables of *ln_dist* and *contig* are time invariant so they drop out automatically when we implement dyad and year fixed effects. This falls in line with our expectations wherein Fixed-effects are useful when you are only interested in analyzing the impact of variables that vary over time. As FE explore the relationship between predictor and outcome variables within an entity (country, person, company, etc.). Here, we are looking at variation within dyads.

Task 1.4

Add countryA-year and countryB-year fixed-effects, modifying the model specification if necessary. Explain the results.

Solution 1.4

Inclusion of the countryA-year and countryB-year fixed effects means that now I'm interested in investigating the variation within a country in a specific year. This means that almost all our control variables (such as GDP, for example) would drop out and be unnecessary as the GDP of a country in a specific year is controlled for by the country-year fixed effect. Similarly, other variables that are time invariant such as distance are also captured by the inclusion of the country fixed effect. As such, now I modify the model specification in the code chunk below to only include the the depth variable and the gattwto variable, as these are variables that are not captured by the fixed effect.

```
## Create country-year variables for Country A & B

gravity_data$countryyear1 <- NA
gravity_data$countryyear2 <- NA
gravity_data$countryyear1 <- paste("cow1", gravity_data$cow1,
                                   gravity_data$year,
                                   sep = "_")

gravity_data$countryyear2 <- paste("cow2",
                                   gravity_data$cow2,
                                   gravity_data$year,
```

```

    sep = "_")

gravity_data$countryyear1 <- as.factor(gravity_data$countryyear1)
gravity_data$countryyear2 <- as.factor(gravity_data$countryyear2)

gravity_data <- gravity_data %>%
  dplyr::select(cow1, cow2, year,
                countryyear1, countryyear2, everything())

### Running the models with the different indexes

model_10 <- felm(data = gravity_data, ln_M ~ depth_rasch_new + gattwto |
  id + countryyear1 + countryyear2)

model_11 <- felm(data = gravity_data, ln_M ~ flex_rasch_new + gattwto |
  id + countryyear1 + countryyear2)

model_12 <- felm(data = gravity_data, ln_M ~ enf_rasch_new + gattwto |
  id + countryyear1 + countryyear2)

## note that in the models above I dropped year cause it doesnt make a difference in the model

screenreg(list(model_10, model_11, model_12))

##
## =====
##               Model 1           Model 2           Model 3
## -----
## depth_rasch_new           0.17 ***
##                        (0.00)
## gattwto                   0.08 ***           0.07 ***           0.08 ***
##                        (0.00)           (0.00)           (0.00)
## flex_rasch_new                                0.23 ***
##                                (0.00)
## enf_rasch_new                                      0.19 ***
##                                      (0.00)
## -----
## Num. obs.             1289722           1289722           1289722
## R^2 (full model)             0.88             0.89             0.89
## R^2 (proj model)             0.01             0.01             0.01
## Adj. R^2 (full model)        0.88             0.88             0.88
## Adj. R^2 (proj model)       -0.04             -0.03             -0.03
## =====
## *** p < 0.001, ** p < 0.01, * p < 0.05

## Let's compare all the different models we have used so far

table_4 <- stargazer(model_1, model_4, model_7, model_10,
  header = FALSE,
  title = "Table 4: Linear Regression by Fixed Effects Type",
  keep = c("depth_rasch_new", "ln_newgdp_67_A", "ln_newgdp_67_B",
    "polity2_A", "polity2_B", "gattwto", "ln_dist", "contig"),
  dep.var.caption = "Dependent Variable: Log Imports (ln_M)",

```

```

model.names = FALSE,
column.labels = c("Pooled OLS", "+Dyad and Year FE",
                  "+Country, Year FE",
                  "+Country-Year FE"),
dep.var.labels.include = FALSE,
font.size = "scriptsize",
model.numbers = FALSE, type = "text"); table_4

```

```

##
## Table 4: Linear Regression by Fixed Effects Type
## =====
##                               Dependent Variable: Log Imports (ln
##                               -----
##                               Pooled OLS      +Dyad and Year FE      +Country, Year FE      +Country-Year FE
## -----
## depth_rasch_new              0.264***      0.173***      0.408***      0.408***
##                               (0.003)      (0.002)      (0.003)      (0.003)
## ln_newgdp_67_A               0.476***      0.485***      0.423***
##                               (0.001)      (0.003)      (0.004)
## ln_newgdp_67_B               0.549***      0.501***      0.437***
##                               (0.001)      (0.002)      (0.003)
## polity2_A                    0.016***      0.004***      0.003***
##                               (0.0002)      (0.0003)      (0.0004)
## polity2_B                    0.018***      -0.002***      -0.003***
##                               (0.0002)      (0.0003)      (0.0004)
## gattwto                      0.147***      0.055***      0.147***
##                               (0.003)      (0.003)      (0.004)
## ln_dist                      -0.621***
##                               (0.002)
## contig                       0.557***
##                               (0.010)
## -----
## Observations                 771,688      828,887      828,887      1,288,887
## R2                           0.634      0.848      0.637      0.637
## Adjusted R2                  0.634      0.844      0.637      0.637
## Residual Std. Error          1.259 (df = 771679)  0.823 (df = 804079)  1.253 (df = 828483)  0.643 (df = 1288887)
## F Statistic                   166,994.300*** (df = 8; 771679)
## =====
## Note:                                                                    *p<0.1; **p<0.01; ***p<0.001
## [1] ""
## [2] "Table 4: Linear Regression by Fixed Effects Type"
## [3] "=====
## [4] "                               Dependent Variable: Log Imports (ln
## [5] "                               -----
## [6] "                               Pooled OLS      +Dyad and Year FE      +Country, Year FE      +Country-Year FE

```



```

## [7] "-----"
## [8] "depth_rasch_new          0.264***          0.173***          0.408***
## [9] "                        (0.003)          (0.002)          (0.003)
## [10] "
## [11] "ln_newgdp_67_A          0.476***          0.485***          0.423***
## [12] "                        (0.001)          (0.003)          (0.004)
## [13] "
## [14] "ln_newgdp_67_B          0.549***          0.501***          0.437***
## [15] "                        (0.001)          (0.002)          (0.003)
## [16] "
## [17] "polity2_A               0.016***          0.004***          0.003***
## [18] "                        (0.0002)         (0.0003)         (0.0004)
## [19] "
## [20] "polity2_B               0.018***          -0.002***         -0.003***
## [21] "                        (0.0002)         (0.0003)         (0.0004)
## [22] "
## [23] "gattwto                 0.147***          0.055***          0.147***
## [24] "                        (0.003)          (0.003)          (0.004)
## [25] "
## [26] "ln_dist                 -0.621***
## [27] "                        (0.002)
## [28] "
## [29] "contig                  0.557***
## [30] "                        (0.010)
## [31] "
## [32] "-----"
## [33] "Observations            771,688            828,887            828,887
## [34] "R2                      0.634              0.848              0.637
## [35] "Adjusted R2             0.634              0.844              0.637
## [36] "Residual Std. Error    1.259 (df = 771679)  0.823 (df = 804079) 1.253 (df = 828483) 0.
## [37] "F Statistic            166,994.300*** (df = 8; 771679)
## [38] "=====
## [39] "Note:

```

*p<0.1; **

Sure enough, the results from Table 4 which shows a comparison of all the different modelling techniques show that our results for the country-year fixed effects were correct. The variable *gattwto* which corresponds to the year in which countries signed the General Agreement on Trade and Tariffs or WTO) and depth (*depth_rasch_new*) are both coming in as positive and statistically significant. While the coefficient on *gattwto* is small (0.08) it still makes sense to report and include it in our model. Moreover, with regards to to depth variable, we see that estimation using country-year fixed effects causes the coefficient on depth to fall further (it is now 0.169 or 0.17). Table 4 above showcases that, with the exclusion of country, year FEs, the coefficient on depth consistently decreases after our initial estimates from the pooled ols model.

There is reason to suspect that there is perhaps omitted variable bias that is positively correlation with depth (and other latent trait variables) that causes us to overestimate the coefficients for our main explanatory variables. The results for the main explanatory variables, *depth*, *flexibility* and *enforcement* remained positive and statistically significant. However, the coefficient of these variables decreases. The exception to this trend are the country, year fixed effects which cause the coefficient on depth to increase even more than the coefficient on depth obtained through a pooled OLS model. Note that the coefficient on depth decreases.

Note that this phenomena is discussed well by Baeier and Bergstrand (2007). They note that a crucial investigation of their paper is how “unobserved heterogeneity in trade flow determinants associated with the likelihood of an FTA?”. More comprehensively, they note that the error term in the gravity model may be representing unobservable (to the econometrician) policy-related barriers — tending to reduce trade between two countries—that are not accounted for by standard gravity equation RHS variables but may be correlated

with the decision to form an FTA (Baeier and Bergstrand, 2007 pg 78). They elucidate this further by giving the example that two countries may have extensive unmeasurable domestic regulations (e.g., internal shipping regulations) that inhibit trade (causing errors in i and j to be negative). The likelihood of the two countries' governments selecting into an FTA may be high if there is a large expected welfare gain from potential bilateral trade creation if the FTA deepens liberalization beyond tariff barriers into domestic regulations (and other non-tariff barriers). Thus, FTA_{ij} and the intensity of domestic regulations may be positively correlated in a cross-section of data, but the gravity equation error term and the intensity of domestic regulations may be negatively correlated. This reason suggests that FTA and error are negatively correlated, and the FTA coefficient will tend to be underestimate.

In other words, for Baeir and Bergstrand their omitted variable is negative correlated with trade but still positively correlated with FTA. As a consequence, their FTA coefficient is underestimated. However, in our case the coefficient on depth (which is like FTA in our case) is *over-estimated* by the previous models. This is due to the fact that the omitted variable is likely to be positively correlated with depth and the outcome variable.

Task 1.5

Run within-estimator dyad fixed effects and compare the results with the ones obtained in point 3 (without year fixed effects).

Solution 1.5

In the following code chunk I first estimate the model using dyad fixed effects through the within estimator approach. Following that, I then reestimate the Least Squares Dummy Variable (LSDV) model for the dyads without the incursion of the year dummies.

```
gravity_data_within <- gravity_data %>%
  group_by(id) %>%
  mutate(
    ln_M = ln_M - mean(ln_M, na.rm = TRUE),
    depth_rasch_new = depth_rasch_new - mean(depth_rasch_new, na.rm = TRUE),
    ln_newgdp_67_A = ln_newgdp_67_A - mean(ln_newgdp_67_A, na.rm = TRUE),
    ln_newgdp_67_B = ln_newgdp_67_B - mean(ln_newgdp_67_B, na.rm = TRUE),
    polity2_A = polity2_A - mean(polity2_A, na.rm = TRUE),
    polity2_B = polity2_B - mean(polity2_B, na.rm = TRUE),
    gattwto = gattwto - mean(gattwto, na.rm = TRUE)
  )

model_13 <- lm(ln_M ~ depth_rasch_new + ln_newgdp_67_A +
  ln_newgdp_67_B + polity2_A + polity2_B +
  gattwto + ln_dist + contig,
  data = gravity_data_within)

model_14 <- lm(ln_M ~ enf_rasch_new +
  ln_newgdp_67_A + ln_newgdp_67_B +
  polity2_A + polity2_B + gattwto +
  ln_dist + contig, data = gravity_data_within)

model_15 <- lm(ln_M ~ flex_rasch_new +
  ln_newgdp_67_A + ln_newgdp_67_B +
  polity2_A + polity2_B + gattwto +
  ln_dist + contig, data = gravity_data_within)

## Let's calculate the dyad fixed effects (without the year)
```

```

model_16 <- felm (data = gravity_data, ln_M ~ depth_rasch_new +
  ln_newgdp_67_A + ln_newgdp_67_B + polity2_A +
  polity2_B + gattwto | id)

model_17 <- felm (data = gravity_data, ln_M ~ flex_rasch_new +
  ln_newgdp_67_A + ln_newgdp_67_B + polity2_A +
  polity2_B + gattwto | id)

model_18 <- felm (data = gravity_data, ln_M ~ enf_rasch_new +
  ln_newgdp_67_A + ln_newgdp_67_B + polity2_A +
  polity2_B + gattwto | id)

## comparison between within-estimator dyad fixed effects and LSDV dyad fixed effects (wihout the year)

table_5 <- stargazer(model_13, model_16,
  header = FALSE,
  title = "Table 5: Within estimator dyad FEs vs Dyad FEs",
  keep = c("depth_rasch_new", "ln_newgdp_67_A", "ln_newgdp_67_B",
    "polity2_A", "polity2_B", "gattwto", "ln_dist", "contig"),
  dep.var.caption = "Dependent Variable: Log Imports (ln_M)",
  model.names = FALSE,
  column.labels = c("Within Estimator Dyad FE", "LSDV - Dyad FEs"),
  dep.var.labels.include = FALSE,
  font.size = "scriptsize",
  model.numbers = FALSE, type = "text"); table_5

```

```

##
## Table 5: Within estimator dyad FEs vs Dyad FEs
## =====
##                               Dependent Variable: Log Imports (ln
##                               -----
##                               Within Estimator Dyad FE      LSDV - Dyad FEs
## -----
## depth_rasch_new          0.237***          0.227***
##                          (0.002)           (0.002)
##
## ln_newgdp_67_A           0.485***          0.563***
##                          (0.002)           (0.002)
##
## ln_newgdp_67_B           0.540***          0.571***
##                          (0.002)           (0.002)
##
## polity2_A                 0.010***          0.015***
##                          (0.0002)          (0.0002)
##
## polity2_B                 0.010***          0.009***
##                          (0.0002)          (0.0002)
##
## gattwto                   0.158***          0.171***
##                          (0.003)           (0.003)
##
## ln_dist                   -0.016***
##                          (0.001)

```

```

##
## contig          0.031***
##                (0.007)
##
## -----
## Observations      771,688          828,887
## R2                0.369          0.844
## Adjusted R2       0.369          0.839
## Residual Std. Error 0.857 (df = 771679) 0.835 (df = 804140)
## F Statistic      56,406.850*** (df = 8; 771679)
## =====
## Note:                *p<0.1; **p<0.05; ***p<0.01

## [1] ""
## [2] "Table 5: Within estimator dyad FEs vs Dyad FEs"
## [3] "=====
## [4] "                Dependent Variable: Log Imports (ln
## [5] "                -----
## [6] "                Within Estimator Dyad FE          LSDV - Dyad FEs
## [7] "-----
## [8] "depth_rasch_new          0.237***          0.227***
## [9] "                (0.002)          (0.002)
## [10] "
## [11] "ln_newgdp_67_A          0.485***          0.563***
## [12] "                (0.002)          (0.002)
## [13] "
## [14] "ln_newgdp_67_B          0.540***          0.571***
## [15] "                (0.002)          (0.002)
## [16] "
## [17] "polity2_A          0.010***          0.015***
## [18] "                (0.0002)          (0.0002)
## [19] "
## [20] "polity2_B          0.010***          0.009***
## [21] "                (0.0002)          (0.0002)
## [22] "
## [23] "gattwto          0.158***          0.171***
## [24] "                (0.003)          (0.003)
## [25] "
## [26] "ln_dist          -0.016***
## [27] "                (0.001)
## [28] "
## [29] "contig          0.031***
## [30] "                (0.007)
## [31] "
## [32] "-----
## [33] "Observations      771,688          828,887
## [34] "R2                0.369          0.844
## [35] "Adjusted R2       0.369          0.839
## [36] "Residual Std. Error 0.857 (df = 771679) 0.835 (df = 804140)"
## [37] "F Statistic      56,406.850*** (df = 8; 771679)
## [38] "=====
## [39] "Note:                *p<0.1; **p<0.05; ***p<0.01"

```

Theoretically, within estimator and the LSDV model should be identical. Table 5 above showcases that while the results that we obtain are not perfectly identical, they are similar enough for us to find evidence in

support of the argument that the results from the within estimation and LSDV model are the same. This is supportive of our theoretical expectations. Note that all the covariates in both models are coming in as statistically significant.

Note however, that the r squared and the adjusted r-squared is very low for the within model (0.369 and 0.369) than the one for the LSDV with dyad FEs model (0.844 and 0.839). The coefficients on the covariates all follow the same directions and are very similar in magnitude. Notably, the coefficient on our main explanatory variable depth is (0.237) for the within model and (0.227) for the LSDV model. In theory, these coefficients should be identical. In practice, they are not too different from each other.

Task 1.6

6. Run a first-difference estimator and explain why results differ from the model with dyad fixed-effects.

Solution 1.6

Here, the question specified to compare the first difference estimations with the results from Question 3. However, as question 1.5 explicitly identified that it was referring to compare with dyad fixed effects without the year, in this question I will include the dyad fixed effects with the year FE.

```
gravity_data_within <- gravity_data %>%
  group_by(id) %>%
  mutate(
    ln_M = ln_M - mean(ln_M, na.rm = TRUE),
    depth_rasch_new = depth_rasch_new - mean(depth_rasch_new, na.rm = TRUE),
    ln_newgdp_67_A = ln_newgdp_67_A - mean(ln_newgdp_67_A, na.rm = TRUE),
    ln_newgdp_67_B = ln_newgdp_67_B - mean(ln_newgdp_67_B, na.rm = TRUE),
    polity2_A = polity2_A - mean(polity2_A, na.rm = TRUE),
    polity2_B = polity2_B - mean(polity2_B, na.rm = TRUE),
    gattwto = gattwto - mean(gattwto, na.rm = TRUE)
  )

gravity_data_fd <- gravity_data %>%
  group_by(id) %>%
  arrange(id, year) %>%
  mutate(
    ln_M = ln_M - lag(ln_M),
    depth_rasch_new = depth_rasch_new - lag(depth_rasch_new),
    ln_newgdp_67_A = ln_newgdp_67_A - lag(ln_newgdp_67_A),
    ln_newgdp_67_B = ln_newgdp_67_B - lag(ln_newgdp_67_B),
    polity2_A = polity2_A - lag(polity2_A),
    polity2_B = polity2_B - lag(polity2_B),
    gattwto = gattwto - lag(gattwto)
  )

## Running the first differences with the three models

model_19 <- lm(ln_M ~ depth_rasch_new + ln_newgdp_67_A +
  ln_newgdp_67_B + polity2_A + polity2_B + gattwto,
  data = gravity_data_fd)

model_20 <- lm(ln_M ~ flex_rasch_new + ln_newgdp_67_A +
  ln_newgdp_67_B + polity2_A + polity2_B +
  gattwto, data = gravity_data_fd)
```

```
model_21 <- lm(ln_M ~ enf_rasch_new + ln_newgdp_67_A +
               ln_newgdp_67_B + polity2_A + polity2_B + gattwto,
               data = gravity_data_fd)
```

```
screenreg(list(model_19, model_20, model_21))
```

```
##
## =====
##               Model 1           Model 2           Model 3
## -----
## (Intercept)      0.03 ***          0.03 ***          0.03 ***
##                 (0.00)            (0.00)            (0.00)
## depth_rasch_new   0.01 ***
##                 (0.00)
## ln_newgdp_67_A    0.20 ***          0.20 ***          0.20 ***
##                 (0.00)            (0.00)            (0.00)
## ln_newgdp_67_B    0.14 ***          0.14 ***          0.14 ***
##                 (0.00)            (0.00)            (0.00)
## polity2_A         -0.00            -0.00            -0.00
##                 (0.00)            (0.00)            (0.00)
## polity2_B         -0.00            -0.00            -0.00
##                 (0.00)            (0.00)            (0.00)
## gattwto           0.01 ***          0.01 ***          0.01 ***
##                 (0.00)            (0.00)            (0.00)
## flex_rasch_new
##                 (0.00)
## enf_rasch_new
##                 (0.00)
## -----
## R^2               0.01             0.01             0.01
## Adj. R^2          0.01             0.01             0.01
## Num. obs.         797150           797150           797150
## RMSE              0.49             0.49             0.49
## =====
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

```
table_6 <- stargazer(model_19, model_4,
                      header = FALSE,
                      title = "Table 6: First differences vs Dyad FEs (with year FEs)",
                      keep = c("depth_rasch_new", "ln_newgdp_67_A", "ln_newgdp_67_B",
                                "polity2_A", "polity2_B", "gattwto", "ln_dist", "contig"),
                      dep.var.caption = "Dependent Variable: Log Imports (ln_M)",
                      model.names = FALSE,
                      column.labels = c("Within Estimator Dyad FE", "LSDV - Dyad FEs (with year FEs)",
                                         "LSDV - Dyad FEs (with year FEs)"),
                      dep.var.labels.include = FALSE,
                      font.size = "scriptsize",
                      model.numbers = FALSE, type = "text")
```

```
##
## Table 6: First differences vs Dyad FEs (with year FEs)
## =====
##               Dependent Variable: Log Imports (ln
## -----
##               Within Estimator Dyad FE    LSDV - Dyad FEs (with year FEs
```

```
## -----
## depth_rasch_new          0.013***          0.173***
##                          (0.003)          (0.002)
##
## ln_newgdp_67_A          0.196***          0.485***
##                          (0.004)          (0.003)
##
## ln_newgdp_67_B          0.145***          0.501***
##                          (0.004)          (0.002)
##
## polity2_A               -0.0003          0.004***
##                          (0.0003)          (0.0003)
##
## polity2_B               -0.0001          -0.002***
##                          (0.0003)          (0.0003)
##
## gattwto                 0.013***          0.055***
##                          (0.004)          (0.003)
##
## -----
## Observations            797,150          828,887
## R2                      0.006          0.848
## Adjusted R2             0.006          0.844
## Residual Std. Error    0.492 (df = 797143)    0.823 (df = 804079)
## F Statistic            861.400*** (df = 6; 797143)
## =====
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

```
screenreg(list(model_19, model_4, model_16))
```

```
##
## =====
##                      Model 1      Model 2      Model 3
## -----
## (Intercept)          0.03 ***
##                      (0.00)
## depth_rasch_new      0.01 ***      0.17 ***      0.23 ***
##                      (0.00)      (0.00)      (0.00)
## ln_newgdp_67_A       0.20 ***      0.48 ***      0.56 ***
##                      (0.00)      (0.00)      (0.00)
## ln_newgdp_67_B       0.14 ***      0.50 ***      0.57 ***
##                      (0.00)      (0.00)      (0.00)
## polity2_A            -0.00          0.00 ***      0.01 ***
##                      (0.00)      (0.00)      (0.00)
## polity2_B            -0.00          -0.00 ***      0.01 ***
##                      (0.00)      (0.00)      (0.00)
## gattwto              0.01 ***      0.05 ***      0.17 ***
##                      (0.00)      (0.00)      (0.00)
## -----
## R^2                  0.01
## Adj. R^2             0.01
## Num. obs.            797150      828887      828887
## RMSE                 0.49
## R^2 (full model)          0.85          0.84
## R^2 (proj model)         0.09          0.40
```

```
## Adj. R^2 (full model)          0.84          0.84
## Adj. R^2 (proj model)         0.07          0.38
## =====
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

As Dur et al note, our data supports the common wisdom that PTAs have become deeper in the past years (Dur et al 2014, 11). This means that there is definitely serial correlation within our data/models that needs to be controlled. For example, the serial correlation is that PTA has been growing and that trade levels has been growing however in the first differences. This this case the effect of depth now is very small. Moreover, earlier on in this assignment we have already noted that strict exogeneity of trade agreements (or in other words our case depth) and our dependent variable (trade levels - imports) is problematic. On a more general, note that both First Differences and FE offer us unbiased estimates whenever $cov(x_{it}, \epsilon_{it})$ is not equal to 0, as long as sequential exogeneity holds (Bailey 2016).

Table 6 above provides us some evidence that in fact the strict exogeneity assumption in panel least squares estimation has been violated. One test of this violation is when results from first differences (FDs) are different from those from LSDV or fixed effects. This is true in our case which gives us some intuition to suspect that the exogeneity assumption in our models has been violated and there is serial correlation in our panel set-up. For example, we can observe from Table 6 that the FD yields us a much lower coefficient of depth so much so that it is substantively uninteresting. Moreover, polity is no longer statistically significant in FDs under the 95 percent significance level. Note that when this “strict exogeneity” fails, both FE and FD are biased.

Theoretically, in a set-up where the strict exogeneity assumption has been violated then both estimators (FDs and FEs) are biased, but in different ways. This will cause FE and FD to look very different. Crucially, in such a scenario it is better to remedy the problem using first differences rather than through FE.